VISIT…

LANZAROTE Caliente.COM

УДК 519.6 (075.8)
ББК 22.19
Р98

Рябенький В.С. **Введение в вычислительную математику:** Учеб. пособие. — 2-е изд., исправл. — М.: ФИЗМАТЛИТ, 2000. — 296 с. — ISBN 5-9221-0047-5.

Изложены основные понятия и идеи, используемые для преобразования математических моделей к виду, удобному для изучения с помощью ЭВМ. Изложение ведется на материале вычислительных задач математического анализа, алгебры и дифференциальных уравнений. Впервые в учебной литературе отражен метод разностных потенциалов для численного решения краевых задач математической физики.

Для студентов механико-математических и физических факультетов университетов, МФТИ, МИФИ, политехнических и других вузов.

Табл. 5. Ил. 42. Библиогр. 24 назв.

Рецензенты:
доктор физико-математических наук *Л.А. Чудов*;
кафедра прикладной математической физики МИФИ.

© ФИЗМАТЛИТ, 2000
© В.С. Рябенький, 1994;
с исправлениями 2000

ISBN 5-9221-0047-5

ПРЕДИСЛОВИЕ

Предлагаемая книга соответствует вводному полугодовому курсу вычислительной математики для студентов Московского физико-технического института, различные варианты которого более двадцати лет читает автор. Книга адресована студентам технических вузов различных профилей.

Цель книги — сообщить достаточно стабильные и общие понятия и идеи, относящиеся к преобразованию математических моделей различных прикладных задач к виду, удобному для вычисления их решений с помощью компьютеров. Изучение книги должно давать возможность сравнительно легко доучиваться для работы в разнообразных и быстро развивающихся конкретных областях. Общая методология сообщается и иллюстрируется в книге на материале численных методов математического анализа, алгебры и дифференциальных уравнений, поскольку эти методы лучше развиты, содержат достаточно совершенные алгоритмы и часто применяются в удаленных друг от друга конкретных прикладных задачах.

Наряду с традиционным для учебной литературы по вычислительной математике материалом книга дает представление о модифицированных граничных уравнениях Кальдерона и методе разностных потенциалов для их численного решения, а также содержит изложение локальных формул гладкого восполнения функций.

При изложении метода конечных разностей использована книга С.К. Годунова и В.С. Рябенького “Разностные схемы”.

Структура книги допускает изучение материала с различной глубиной. Это достигается расположением материала и подбором задач.

Для более полного и глубокого изучения многих вопросов, освещенных в учебнике, можно воспользоваться книгами К.И. Бабенко [1], Н.С. Бахвалова, Н.П. Жидкова и Г.М. Кобелькова [2], С.К. Годунова и В.С. Рябенького [6, 7], Н.Н. Калиткина [9], Г.И. Марчука [12], А.А. Самарского [18], А.А. Самарского и Е.С. Николаева [20], а также другими книгами и журнальными публикациями, указанными в тексте.

Отметим еще, что изучение книги можно сделать более наглядным и увлекательным, если воспользоваться предназначенным для этого

циклом компьютерных учебных демонстрационно-лабораторных работ. Этот цикл создан сотрудниками кафедры вычислительной математики МФТИ В.Д. Ивановым, П.Н. Коротиным, В.И. Косаревым, И.Б. Петровым (руководитель работ), В.Б. Пироговым, В.С. Рябеньким (научный руководитель), Д.С. Северовым, А.Г. Тормасовым, С.В. Утюжниковым. В программной реализации цикла участвовали студенты В.В. Бойков, Д.Л. Будько, К.Б. Бухаров, А.Ю. Езерский, А.Б. Константинов, С.А. Корытник, Ю.П. Кравченко, Ю.Д. Крикунов, Д.В. Лунев, В.А. Торгашов, А.А. Терехин, Г.Л. Химичев. Краткое описание цикла приведено в конце книги.

В заключение автор сердечно благодарит академика О.М. Белоцерковского, который в момент создания кафедры вычислительной математики МФТИ поручил ему чтение первого обязательного общего курса вычислительной математики. Олег Михайлович неоднократно обсуждал с автором цели курса в условиях физтеха и соответствующий этим целям характер курса. Подготовка и чтение лекций оказались началом работы автора над предлагаемой книгой.

Автор сердечно благодарит также всех своих коллег по работе в МФТИ и особенно В.В. Демченко, В.Д. Иванова, В.И. Косарева, В.Я. Митницкого, Н.П. Онуфриеву, И.Б. Петрова, В.Б. Пирогова, Л.М. Стрыгину, Р.П. Федоренко, А.С. Холодова и Л.А. Чудова за многолетнее сотрудничество, оказавшее влияние на книгу.

Автор сердечно благодарит своего коллегу К.В. Брушлинского, прочитавшего книгу в рукописи, за ряд полезных замечаний.

Автор выражает глубокую благодарность профессору К.И. Бабенко и профессору О.В. Локуциевскому, с которыми также плодотворно обсуждал вопросы преподавания.

В.С. Рябенький

ВВЕДЕНИЕ

Современная вычислительная математика ориентирована на использование компьютеров для прикладных расчетов. Любые математические приложения начинаются с построения модели явления (изделия, действия, ситуации или другого объекта), к которому относится изучаемый вопрос. Классическими примерами математических моделей могут служить определенный интеграл, уравнение колебаний маятника, уравнение теплообмена, уравнения упругости, уравнения электромагнитных волн и другие уравнения математической физики. Назовем еще для контраста модель формальных рассуждений — алгебру Буля.

Основополагающими средствами изучения математических моделей являются аналитические методы: получение точных решений в частных случаях (например, табличные интегралы), разложения в ряды. Определенную роль издавна играли приближенные вычисления. Например, для вычисления определенного интеграла использовались квадратурные формулы.

Появление в середине XX века электронных вычислительных машин (компьютеров) радикально расширило возможности приложения математических методов в традиционных областях (механике, физике, технике) и вызвало бурное проникновение математических методов в нетрадиционные области (управление, экономику, химию, биологию, психологию, лингвистику, экологию и т. п.).

Компьютер дает возможность запоминать большие (но конечные) массивы чисел и производить над ними арифметические операции и сравнения с большой (но конечной) скоростью по заданной вычислителем программе. Поэтому на компьютере можно изучать только те математические модели, которые описываются конечными наборами чисел, и использовать конечные последовательности арифметических действий, а также сравнений чисел по величине (для автоматического управления дальнейшими вычислениями).

В традиционных областях математическими моделями служат функции, производные, интегралы, дифференциальные уравнения. Для использования компьютеров эти исходные модели надо приближенно заменить такими, которые описываются конечными наборами

чисел с указанием конечных последовательностей действий (конечных алгоритмов) для их обработки. Например, функцию следует заменить таблицей; производную

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

заменить приближенной формулой

$$f'(x) \approx \frac{f(x+h) - f(x)}{h};$$

определенный интеграл — суммой; краевую задачу для дифференциального уравнения — задачей об отыскании таблицы значений решения в узлах некоторой сетки, причем так, чтобы выбор шага сетки позволял достигать любой требуемой точности. Оказывается, из двух, на первый взгляд равноценных способов один может оказаться принципиально непригодным из-за того, что доставляемое им приближенное решение не стремится к искомому при измельчении шага сетки, или из-за катастрофически сильной чувствительности к погрешностям округления.

Теория таких моделей и алгоритмов составляет предмет вычислительной математики. Эта теория тесно связана с теориями приближения и интерполяции функций, уравнений с частными производными, интегральных уравнений, информационной сложности функциональных классов, алгоритмов, а также с языками программирования для расчетов на компьютере и т. д. Современные вычислительные методы позволяют, например, рассчитать характеристики обтекания газом тела заданной формы, что недоступно аналитическим методам (подобно нетабличному интегралу).

С использованием компьютера стал возможен вычислительный эксперимент, т. е. расчет в целях проверки гипотез, а также в целях наблюдения за поведением модели, когда заранее не известно, что именно заинтересует исследователя. В процессе численного эксперимента происходит по существу уточнение исходной математической постановки задачи. В процессе расчетов на компьютере происходит накопление информации, что дает возможность в конечном счете произвести отбор наиболее интересных ситуаций. На этом пути сделано много наблюдений и открытий, стимулирующих развитие теории и имеющих важные практические применения.

С помощью компьютера возможно применение математических методов и в нетрадиционных областях, где не удается построить компактные математические модели вроде дифференциальных уравнений, но удается построить модели, доступные запоминанию и изучению на компьютере. Модели для компьютеров в этих случаях представляют собой цифровое кодирование схемы изучаемого объекта (например, языка) и отношений между его элементами (словами,

фразами). Сама возможность изучения таких моделей на компьютере стимулирует появление этих моделей, а для создания обозримой модели необходимо выявление законов, действующих в исходных объектах. С другой стороны, получаемые на компьютере результаты (например, машинный перевод упрощенных текстов с одного языка на другой) вносят критерий практики в оценку теорий (например, лингвистических теорий), положенных в основу математической модели.

Благодаря компьютерам стало возможным рассматривать вероятностные модели, требующие большого числа пробных расчетов, имитационные модели, которые отражают моделируемые свойства объекта без упрощений (например, функциональные свойства телефонной сети).

Разнообразие задач, где могут быть использованы компьютеры, очень велико. Для решения каждой задачи нужно знать многое, связанное именно с этой задачей. Естественно, этому нельзя научиться впрок.

Целью этой книги является сообщение тех основных понятий, идей и методов, владение которыми позволяет сравнительно быстро научиться работать в конкретных областях. Это реализуется на материале вычислительных задач алгебры, математического анализа, дифференциальных уравнений, поскольку здесь методы хорошо развиты и применяются в далеких друг от друга областях.

Назовем некоторые общие понятия и идеи, которые требуют внимания и наполняются конкретным содержанием в зависимости от задачи, которую предстоит решать с помощью компьютера. Это — дискретизация задачи; обусловленность задачи; погрешность численного метода; вычислительная устойчивость алгоритма; сравнение алгоритмов по полноте используемой ими входной информации, по используемой памяти и числу арифметических действий. Алгоритмы могут отличаться возможностью распараллеливания для одновременного проведения вычислений на многопроцессорном компьютере. Одним из плодотворных и основных методов вычислительной математики является комбинированное использование аналитических и компьютерных средств.

Здесь, во введении, мы предварительно познакомим читателя с перечисленными понятиями. Это даст некоторое общее представление о предмете вычислительной математики и подготовит к изучению дальнейшего материала.

§ 1. Дискретизация

Пусть требуется найти приближенное решение какой-либо задачи, в которой в качестве входных данных участвует какая-либо функция $f(x)$, определенная на всем (бесконечном) множестве точек отрезка

$0 \leqslant x \leqslant 1$. Значения этой функции при каждом фиксированном x можно получить измерениями или вычислениями. Для запоминания этой функции в памяти компьютера необходимо приближенно описать ее таблицей значений на некотором конечном множестве отдельных точек x_0, x_1, ..., x_n. Это — простейший пример дискретизации задачи: от задачи запоминания функции на отрезке $[0, 1]$ мы перешли к задаче запоминания таблицы значений на дискретном множестве точек $x_0, x_1, ..., x_n$ из этого отрезка.

Пусть функция $f(x)$ имеет достаточное число производных, а нам требуется вычислить ее производную $f'(x)$ в данной точке x. Задачу отыскания

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$

содержащую предельный переход, можно заменить приближенно задачами вычисления по одной из формул

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}, \tag{1}$$

$$f'(x) \approx \frac{f(x) - f(x-h)}{h}, \tag{2}$$

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}. \tag{3}$$

Для замены производной $f''(x)$ можно воспользоваться формулой

$$f''(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}. \tag{4}$$

Все эти формулы уточняются при уменьшении h, а при каждом фиксированном h определены для конечных наборов значений функции и используют только арифметические операции. Эти формулы — примеры дискретизации задачи о вычислении производных $f'(x)$, $f''(x)$.

Рассмотрим краевую задачу

$$\begin{gathered} \frac{d^2y}{dx^2} - x^2 y = \cos x, \quad 0 \leqslant x \leqslant 1, \\ y(0) = 2, \quad y(1) = 3, \end{gathered} \tag{5}$$

об отыскании функции $y(x)$, определенной на отрезке $0 \leqslant x \leqslant 1$. Для построения приближенной дискретной модели этой задачи осуществим следующие два шага.

Разобьем отрезок $0 \leqslant x \leqslant 1$ на N равных частей длины $h = N^{-1}$ каждая, а вместо функции $y(x)$ будем искать набор значений y_0, y_1, ..., y_N этой функции в точках $x_k = kh$ $(k = 0, 1, ..., N)$. В точках x_k $(k = 1, 2, ..., N - 1)$ заменим производную $y''(x)$ приближенно по формуле (4) и получим

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - (kh)^2 y_k = \cos kh, \quad k = 1, 2, ..., N-1. \tag{6}$$

Кроме того, в силу граничных условий (5) положим

$$y_0 = 2, \quad y_N = 3. \tag{7}$$

Система $N+1$ линейных уравнений (6), (7) относительно того же числа неизвестных y_0, y_1, ..., y_N является дискретным аналогом задачи (5).

Есть основания думать, что с ростом N решение задачи (6), (7) есть все более точная таблица значений решения задачи (5) (в дальнейшем это будет показано).

Обозначим континуальную краевую задачу через M_∞, а дискретную краевую задачу (6), (7) через M_N. Тогда можно сказать, что задаче M_∞ мы сопоставили бесконечную последовательность дискретных задач M_N $(N = 2, 3, \ldots)$.

Вычисляя решение задачи M_N при каком-либо фиксированном N, мы имеем дело с конечным набором чисел, задающих входные данные, и с конечным набором чисел y_0, y_1, ..., y_N, подлежащих отысканию. Однако вычислительная математика обычно ставит своей целью предложить именно последовательность уточняющихся дискретных моделей M_N, так как это дает возможность выбрать то N, которое обеспечивает выполнение требований к точности.

Переход от континуальной задачи M_∞ к последовательности M_N ее дискретных моделей возможен многими способами. Пусть $\{M_N\}$, $\{M'_N\}$ — какие-нибудь две последовательности таких моделей, причем вычисление решений дискретных задач M_N, M'_N требует равных затрат. Тогда предпочтение надо отдать тому способу дискретизации, при котором решение дискретной задачи может служить решением исходной задачи с заданной точностью при меньшем значении N.

Бывает, что из двух, казалось бы, равноценных способов дискретизации M_N и M'_N один при возрастании N дает все более точное приближение к решению континуальной задачи M_∞, а другой приводит к "приближенному решению" задачи, которое с ростом N теряет какое-либо сходство с искомым решением. С этой ситуацией и способами ее преодоления мы встретимся в части III книги.

Задача

Пусть $f(x)$ имеет ограниченные производные до требуемого порядка. Показать, что погрешности приближенных формул (1)–(4) суть величины $O(h)$, $O(h)$, $O(h^2)$, $O(h^2)$.

§ 2. Обусловленность

Во всякой задаче требуется по входным данным сделать заключение о каких-либо свойствах решения. Похожие на первый взгляд задачи могут резко отличаться чувствительностью интересующих

свойств решения к возмущению входных данных. Если эта чувствительность “мала”, то задача считается хорошо обусловленной; в противном случае — плохо обусловленной. Обычно плохо обусловленные задачи не только предъявляют высокие требования к точности задания входных данных, но и более трудны для вычислений.

Пример. Пусть концентрация $y = y(t)$ некоторого вещества в момент времени t есть функция, удовлетворяющая дифференциальному уравнению

$$\frac{dy}{dt} - 10y = 0.$$

Фиксируем произвольно t_0 $(0 \leqslant t_0 \leqslant 1)$ и делаем приближенное измерение y_0^* концентрации $y_0 = y(t_0)$, получив

$$y\big|_{t=t_0} = y_0^*.$$

Задача состоит в определении концентрации $y = y(t)$ в произвольный момент времени t из отрезка $0 \leqslant t \leqslant 1$.

Если бы число $y_0 = y(t_0)$ было известно точно, то можно было бы указать точную формулу

$$y(t) = y_0 e^{t-t_0}$$

для концентрации. Но мы знаем лишь приближенное значение $y_0^* \approx y_0$ числа y_0. Поэтому вместо $y(t) = y_0 e^{10(t-t_0)}$ мы можем указать лишь приближенную формулу $y^*(t) = y_0^* e^{10(t-t_0)}$. Очевидно, что погрешность $y^* - y$ выражается формулой

$$y^*(t) - y(t) = (y_0^* - y_0)e^{10(t-t_0)}, \quad 0 \leqslant t \leqslant 1.$$

Допустим, нам нужно произвести замер y_0^* с такой точностью δ, $|y_0^* - y_0| < \delta$, чтобы гарантировать некоторую заданную точность $\varepsilon > 0$ всюду на отрезке $0 \leqslant t \leqslant 1$, т. е. гарантировать оценку

$$|y^*(t) - y(t)| < \varepsilon, \quad 0 \leqslant t \leqslant 1.$$

Очевидно, $\max\limits_{0 \leqslant t \leqslant 1} |y^*(t) - y(t)| = |y^*(1) - y(1)| = |y_0^* - y_0|e^{10(1-t_0)}$.

Отсюда получаем следующее требование к точности δ измерения y_0:

$$\delta \leqslant \varepsilon e^{-10(1-t_0)}.$$

Пусть измерение y_0 производится в момент $t_0 = 0$. Тогда требование к точности δ измерения будет в e^{10} раз, т. е. в тысячи раз, выше, чем требуемая гарантированная точность ε результата. Ответ весьма чувствителен к погрешности задания входных данных, т. е. y_0, и задача плохо обусловлена.

Если измерение производить при $t_0 = 1$, то $\delta = \varepsilon$, т. е. достаточно измерения с гораздо меньшей точностью, чем в случае $t_0 = 0$, и задача хорошо обусловлена.

Задачи

1. На каком из двух отрезков, $x[1/2, 1]$ или $[-1, 0]$, задача вычисления $y(1+x)/(1-x)$ по x лучше обусловлена.

2. Пусть $y = \sqrt{2} - 1$. Можно написать также $y = (\sqrt{2}+1)^{-1}$. Какая из двух формул чувствительнее к погрешности при приближенном задании $\sqrt{2}$ в виде конечной десятичной дроби?

Указание. Сравнить модули производных функций $(x-1)$ и $(x+1)^{-1}$.

§ 3. Погрешность

Во всякой вычислительной задаче по некоторым входным данным задачи требуется найти ответ на поставленный вопрос. Если ответ на вопрос задачи можно дать с абсолютной точностью, то погрешность отсутствует. Но обычно удается найти ответ лишь с некоторой погрешностью. Погрешность вызывается тремя причинами.

Первая причина — некоторая неопределенность при задании входных данных, которая приведет к соответствующей неопределенности в ответе: ответ может быть указан лишь с некоторой погрешностью, которая носит название *неустранимой погрешности.*

Вторая причина: если мы ликвидируем неопределенность в задании входных данных, фиксировав какие-либо входные данные, а затем будем вычислять ответ с помощью какого-нибудь приближенного метода, то найдем не в точности тот ответ, который соответствует этим фиксированным входным данным. Возникает погрешность, связанная с выбором приближенного метода вычислений.

Третья причина: сам выбранный нами приближенный метод реализуется неточно из-за погрешностей округлений при вычислениях на реальном компьютере.

Погрешность результата складывается, таким образом, из неустранимой погрешности, погрешности метода и погрешности округлений.

Проиллюстрируем эти понятия.

1. *Неустранимая погрешность.* Пусть задача состоит в вычислении значения y некоторой функции $y = f(x)$ при некотором $x = t$. Число t и соответствие f, в силу которого каждому значению аргумента сопоставляется значение функции, служат входными данными задачи, а число $y(t)$ — решением.

Пусть функция $f(x)$ известна лишь приближенно, например, $f(x) \approx \approx \sin x$, причем известно лишь, что $f(x)$ отличается от $\sin x$ не более, чем на некоторую величину $\varepsilon > 0$:

$$\sin x - \varepsilon \leqslant f(x) \leqslant \sin x + \varepsilon. \tag{1}$$

Пусть значение аргумента $x = t$ получается приближенным измерением, в результате которого получаем некоторое $x = t^*$, причем известно

лишь, что t лежит в пределах

$$t^* - \delta \leqslant t \leqslant t^* + \delta, \tag{2}$$

где $\delta > 0$ — число, характеризующее точность измерения.

Из рис. 1 видно, что величиной $y = f(t)$ может оказаться любая точка отрезка $[a, b]$, где $a = \sin(t^* - \delta) - \varepsilon$, $b = \sin(t^* + \delta) + \varepsilon$. Понятно,

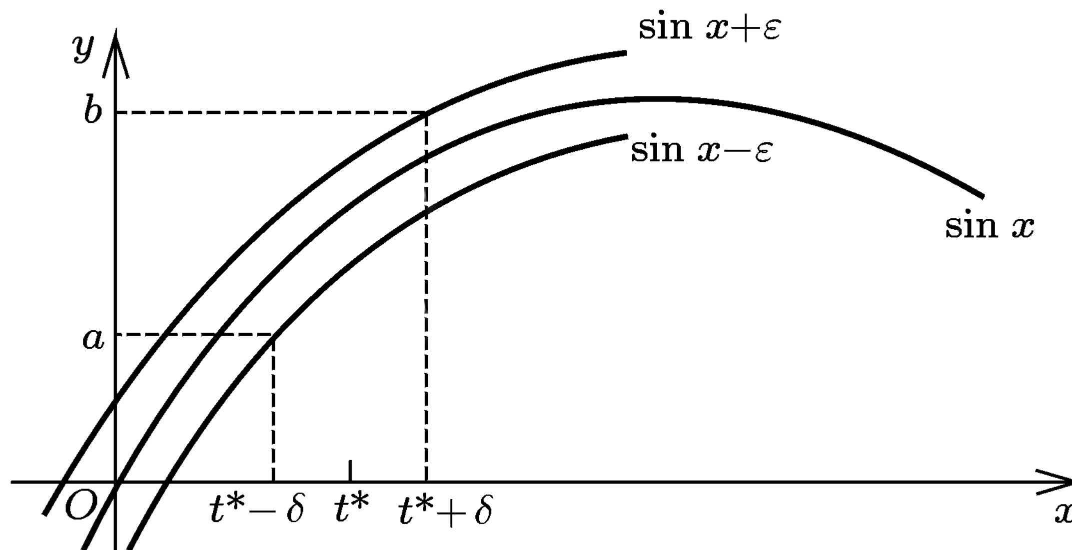


Рис. 1

что, приняв за приближенное значение числа $y = f(t)$ любую точку y^* отрезка $[a, b]$, можно гарантировать оценку погрешности

$$|y - y^*| \leqslant |b - a|. \tag{3}$$

Эту оценку погрешности нельзя существенно уменьшить при имеющихся неполных входных данных. Самая малая погрешность, которую можно гарантировать, получается, если принять за y^* середину отрезка $[a, b]$, положив

$$y^* = y^*_{\text{опт.}} = (a + b)/2.$$

Из рис. 1 видно, что гарантирована оценка

$$|y - y^*_{\text{опт.}}| \leqslant |b - a|/2. \tag{4}$$

Это неравенство станет точным равенством, если $y(t) = a$ или $y(t) = b$.

Таким образом, $|b - a|/2$ и есть та неустранимая (неуменьшаемая) погрешность, которую можно гарантировать при имеющихся неопределенных входных данных в случае самого удачного выбора приближенного решения $y^*_{\text{опт.}}$.

Оптимальная оценка (4) ненамного лучше оценки (3). Поэтому мы не отступим от здравого смысла, если о любой точке $y^* \in [a, b]$, а не только о точке $y^*_{\text{опт.}}$, условимся говорить, что она есть приближенное решение задачи вычисления числа $y(t)$, найденное с неустранимой погрешностью, а вместо $|b - a|/2$ из (4) за величину неустранимой погрешности примем (условно) число $|b - a|$.

2. *Погрешность метода.* Положим $y^* = \sin t^*$. Число y^* принадлежит отрезку $[a, b]$ и является неулучшаемым приближенным решением задачи, погрешность которого удовлетворяет оценке (3) и неустранима. Точка $y^* = \sin t^*$ выбрана среди других точек отрезка $[a, b]$,

потому что она задается удобной для дальнейшей работы формулой.

Для вычисления числа $y^* = \sin t^*$ на компьютере воспользуемся разложением функции $\sin x$ в ряд:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots \tag{5}$$

Для вычисления y^* можно воспользоваться одним из приближенных выражений

$$\begin{aligned} &y^* \approx y_1^* = t^*, \\ &y^* \approx y_2^* = t^* - \frac{(t^*)^3}{3!}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &y^* \approx y_n^* = \sum_{k=1}^{n} (-1)^{k-1} \frac{(t^*)^{2k-1}}{(2k-1)!}. \end{aligned} \tag{6}$$

Выбирая для приближенного вычисления y^* одну из формул (6), мы тем самым выбираем метод вычисления.

Величина $|y^* - y_n^*|$ есть *погрешность метода вычислений.*

Фактически выбранный нами метод вычисления зависит от параметра n и позволяет добиваться, чтобы погрешность метода была меньше любой наперед заданной величины за счет выбора этого параметра.

Очевидно, нет смысла добиваться, чтобы погрешность метода была существенно (во много раз) меньше неустранимой погрешности. Число n поэтому нет смысла брать слишком большим. Однако в случае, если n выбрано слишком маленьким так, что погрешность метода существенно больше неустранимой погрешности, то избранный метод не полностью использует информацию о решении, содержащуюся во входных данных, теряя часть этой информации.

3. *Погрешность округлений.* Допустим, мы фиксировали метод вычислений, положив $y^* \approx y_n^*$. При вычислении y_n^* по формуле (6) на реальном компьютере в результате округлений мы получим некоторое число $\tilde{y}_n^*$. Погрешность $|y_n^* - \tilde{y}_n^*|$ будем называть *погрешностью округлений.*

Эта погрешность не должна быть существенно больше погрешности метода вычислений. В противном случае произойдет потеря точности метода за счет погрешностей округлений.

Задачи *)

1. Пусть требуется вычислить значение $y = f(x)$ функции $f(x)$ по неполным входным данным x^* $(x^* - \delta \leqslant x \leqslant x^* + \delta)$.

Какова неустранимая погрешность, вызванная неполным значением входных данных, в зависимости от x^* и $\delta > 0$:

а) $f(x) = \sin x$;

б) $f(x) = \ln x$, $x > 0$?

*) Символом * будут отмечены задачи повышенной сложности.

При каких значениях x^*, полученных приближенным измерением неопределенной величины x с погрешностью δ, в задаче б) можно указать лишь одностороннюю оценку для $\ln x$ сверху. Укажите эту оценку.

2*. Пусть функция $f(t)$ задана таблицей своих значений в точках $t_k = kh$ $(h = 1/N,\ k = 0, \pm 1, \pm 2, \ldots)$. Пусть о функции $f(t)$ известно кроме этой таблицы еще, что $\max\limits_x |f''(x)| \leqslant 1$.

Показать, что неполные входные данные, содержащиеся в таблице, вообще говоря, не позволяют восстановить в произвольной наперед заданной точке t функцию точнее, чем с неустранимой погрешностью $\varepsilon(h) = h^2/\pi^2$.

Указание. Показать, что наряду с функцией $f(t) \equiv 0$, имеющей нулевую таблицу, функция $\varphi(t) = (h^2/\pi^2)\sin(N\pi t)$ также имеет нулевую таблицу и удовлетворяет условию $\max\limits_x |\varphi''(x)| \leqslant 1$, при этом $\max\limits_t |f(t) - \varphi(t)| = h^2/\pi^2$.

3. Задана некоторая функция $y = f(x)$, о которой известно, что ее вторая производная по модулю не превосходит единицы.

Показать, что погрешность приближенного равенства

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

не превосходит величину h.

4. Пусть некоторая функция $y = f(x)$ имеет вторую производную $f''(x)$, которая по модулю не превосходит единицы. При каждом x значение $f(x)$ получается приближенным измерением величины $f(x)$ и оказывается равным некоторому числу $f^*(x)$. Пусть известно лишь, что точность измерения гарантирует справедливость оценки

$$|f(x) - f^*(x)| \leqslant \varepsilon,$$

где $\varepsilon > 0$ — число, характеризующее точность измерений. Пусть требуется приближенно вычислить $f'(x)$.

а) Как выбрать h, чтобы гарантированная погрешность приближенной формулы

$$f'(x) \approx \frac{f^*(x+h) - f^*(x)}{h}$$

была наименьшей?

б) Показать, что неустранимая погрешность задачи вычисления $f'(x)$ при заданных (не вполне определенных) входных данных не меньше $O(\sqrt{\varepsilon})$.

Указание к б). Обе функции, $f(x) \equiv 0$, $f^*(x) = \varepsilon \sin(x/\sqrt{\varepsilon})$, имеют вторые производные, не превосходящие единицы по модулю, и $\max\limits_x |f(x) - f^*(x)| \leqslant \varepsilon$. В то же время

$$\left|\frac{df^*}{dx} - \frac{df}{dx}\right| = \left|\sqrt{\varepsilon}\cos\frac{x}{\sqrt{\varepsilon}}\right| = O(\sqrt{\varepsilon}).$$

Сопоставляя результаты пп. а), б), проверить, что метод вычисления производной из п. а) дает неулучшаемый по порядку погрешности $O(\sqrt{\varepsilon})$ результат, а также показать, что порядок неустранимой погрешности в точности совпадает с $O(\sqrt{\varepsilon})$.

5. Для запоминания сведений о линейной функции $f(x) = kx + b$, $\alpha \leqslant x \leqslant \beta$, удовлетворяющей неравенствам $0 \leqslant f(x) \leqslant 1$, имеется шесть занумерованных клеток, в каждую из которых можно записать одну из десяти цифр: $0, 1, \ldots, 9$.

Какова неустранимая погрешность при восстановлении функции, если эти шесть клеток заполнены одним из указанных здесь способов?

а) В первые три клетки записаны соответствующие первые три цифры, стоящие после запятой в записи $f(\alpha)$ в виде десятичной дроби, а в три оставшиеся клетки — первые три цифры, стоящие после запятой в записи $f(\beta)$ в виде десятичной дроби.

б) Пусть $\alpha = 0$, $\beta = 10^{-2}$. В первые три клетки записаны первые три цифры десятичной записи числа k, в четвертую клетку — 0 или 1 в зависимости от знака числа k, а в оставшиеся две клетки записаны первые две цифры после запятой из десятичной записи числа b.

в*) Показать, что при любом способе задания функции из заданного класса с помощью шестиместной таблицы указанного вида неустранимая погрешность не меньше $0{,}5 \cdot 10^{-3}$.

Указание. Построить 10^6 функций из заданного класса, максимум модуля разности любых двух из которых не меньше 10^{-3}.

§ 4. О методах вычисления

Пусть для изучения некоторого объекта построена его математическая модель, которая и подлежит изучению средствами вычислительной математики.

Например, математической моделью малых колебаний маятника может оказаться следующая задача:

$$\begin{gathered} \frac{d^2y}{dt^2} + y = 0, \quad t \geqslant 0, \\ y(0) = 0, \quad \left.\frac{dy}{dt}\right|_{t=0} = 1, \end{gathered} \tag{1}$$

где $y(t)$ — отклонение маятника в момент времени t.

При изучении гармонических колебаний с помощью этой математической модели, т. е. задачи Коши (1), некоторую пользу может принести знание физической сущности моделируемого физического объекта. В данной задаче, например, можно предвидеть, что решение носит колебательный (периодический) характер. Однако математическая модель (1) после ее построения становится самостоятельным объектом, для изучения которого можно применять любые математические средства, в том числе и те, которые не имеют никакого отношения к физическому происхождению задачи. Это сильно расширяет возможности исследователя. Так, например, значение решения $y = \sin t$ задачи (1) в заданный момент времени $t = z$, т. е. представление числа $\sin z$ в виде десятичной дроби с заданным числом знаков, можно осуществить с помощью ряда

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \ldots,$$

воспользовавшись его частичной суммой.

Между тем представление функции $\sin t$ в виде степенного ряда, использованное нами, не имеет никакого физического смысла.

Для численного решения какой-либо задачи на компьютере возможны различные численные методы, или различные алгоритмы. Одни из них могут быть много лучше других.

В этой книге мы изложим хорошие алгоритмы для многих часто встречающихся классов задач, а пока объясним, чем могут различаться алгоритмы.

Пусть для вычисления решения y некоторой задачи по входным данным, совокупность которых обозначим X, имеются алгоритмы A_1, A_2, которые дают приближения $y_1^* = A_1(X)$, $y_2^* = A_2(X)$. При этом может оказаться следующее.

1. Алгоритм A_2 может быть точнее алгоритма A_1, т. е.

$$|y - y_1^*| \gg |y - y_2^*|.$$

Например, будем вычислять $y = \sin x|_{x=0,1}$ по формуле вида

$$y_n^* = \sum_{k=1}^{n} (-1)^{k-1} \frac{x^{2k-1}}{(2k-1)!}. \tag{2}$$

Примем за A_1 алгоритм, отвечающий $n = 1$, а за A_2 алгоритм, отвечающий $n = 2$. Очевидно, что

$$|\sin 0{,}1 - y_1^*| \gg |\sin 0{,}1 - y_2^*|.$$

2. Алгоритмы обладают одинаковой точностью, но вычисление $y_1^* = A_1(X)$ требует много больше арифметических действий, чем вычисление y_2^*.

Пусть, например, требуется найти

$$y = 1 + x + x^2 + \ldots + x^{1023} \qquad \left(y = \frac{1 - x^{1024}}{1 - x}\right)$$

при $x = 0{,}99$. За A_1 примем алгоритм, осуществляющий вычисления прямо по заданной формуле, возводя 0,99 поочередно в степени $1, 2, \ldots, 1023$ и складывая. В качестве A_2 примем алгоритм, осуществляющий вычисление по формуле

$$y = \frac{1 - 0{,}99^{1024}}{1 - 0{,}99}.$$

По точности алгоритмы совпадают (оба абсолютно точны), но первый требует гораздо больше арифметических операций.

А именно, вычисляя последовательно

$$x, \quad x^2 = x \cdot x, \quad \ldots, \quad x^{1023} = x^{1022} \cdot x,$$

придется проделать 1022 умножения. В то же время при вычислении $0{,}99^{1024}$ требуется всего 10 умножений:

$$0{,}99^2 = 0{,}99 \cdot 0{,}99, \quad 0{,}99^4 = (0{,}99^2) \cdot (0{,}99^2), \quad \ldots$$
$$\ldots, \quad 0{,}99^{1024} = (0{,}99^{512}) \cdot (0{,}99^{512}).$$

3. Алгоритмы обладают одинаковой точностью, но $A_1(X)$ вычислительно устойчив, а $A_2(X)$ неустойчив.

Например, для вычисления $y = \sin x$ с заданной точностью $\varepsilon = 10^{-3}$ $(|y - y^*| \leqslant 10^{-3})$ воспользуемся суммой

$$y_1^* = y_1^*(x) = \sum_{k=1}^{n} (-1)^{k-1} \frac{x^{2k-1}}{(2k-1)!}, \tag{3}$$

где $n = n(\varepsilon)$ выбирается так, чтобы выполнялось неравенство

$$|y - y_1^*| \leqslant 10^{-3},$$

приняв вычисление этой суммы за алгоритм A_1. Если $|x| \leqslant \pi/2$, то

$$\frac{1}{(2n-1)!}\left(\frac{\pi}{2}\right)^{2n-1} \leqslant 10^{-3}$$

при $n = 5$, и

$$y_1^* = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}.$$

Очевидно, что вычисления по этой формуле слабо реагируют на погрешности округления при вычислении каждого из слагаемых, а поэтому алгоритм A_1 в данном случае вычислительно устойчив.

Пусть $|x| \gg 1$; например, $x = 100$. Тогда для достижения требуемой точности число n должно удовлетворять неравенству

$$\frac{100^{2n+1}}{(2n+1)!} \leqslant 10^{-3},$$

из которого для n заведомо выполнено неравенство $n > 48$. Но при вычислении суммы (3) первые ее члены будут очень большими. Малая относительная погрешность при их вычислении дает большую абсолютную погрешность, и алгоритм A_1 вычислительно неустойчив.

Укажем устойчивый алгоритм A_2. Записываем заданное x в виде $x = l\pi + z$ $(|z| \leqslant \pi/2,$ l целое). Тогда

$$\sin x = (-1)^l \sin z,$$

$$y^* = A_2(x) = (-1)^l\left(z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!}\right).$$

Этот алгоритм по устойчивости совпадает с алгоритмом A_1 для $|x| \leqslant \pi/2$.

4. Алгоритм может быть сходящимся или расходящимся.

Пусть требуется вычислить $y = \ln(1 + x)$. Воспользуемся рядом

$$y = \ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots \tag{4}$$

и положим

$$y^*(x) \approx y_n^* = \sum_{k=1}^{n} (-1)^{k+1} \frac{x^k}{k}.$$

Получим метод вычисления $y = \ln(1+x)$, зависящий от номера n как от параметра.

Если $|x| = q < 1$, то $\lim\limits_{n\to\infty} y_n^*(x) = y(x)$, т. е. погрешность алгоритма с ростом n стремится к нулю. Но если $x > 1$, то $\lim\limits_{n\to\infty} y_n^*(x) = \infty$, так как ряд (4) имеет радиус сходимости $r = 1$. В этом случае алгоритм непригоден для вычислений.

В книге мы познакомимся и с некоторыми другими характеристиками алгоритмов. Встретятся алгоритмы, допускающие параллельные вычисления или требующие последовательного выполнения операций, самонастраивающиеся на специфику входных данных или учитывающие ее лишь отчасти, алгоритмы логически простые или более сложные.

Задачи

1. Предложить алгоритм вычисления $y = \ln(1+x)$, пригодный при $x > 1$.

2. Рассмотрим задачу об определении последовательности $x_0, x_1, ..., x_N$, удовлетворяющей уравнению

$$2x_n - x_{n+1} = 1 + n^2/N^2, \quad n = 0, 1, ..., N-1,$$

и дополнительному условию

$$x_0 + x_N = 1. \tag{5}$$

Предлагаются следующие два алгоритма. Полагаем

$$x_n = u_n + cv_n, \quad n = 0, 1, ..., N.$$

В алгоритме A_1 определяем u_n $(n = 0, 1, ..., N)$ как решение уравнения

$$2u_n - u_{n+1} = 1 + n^2/N^2, \quad n = 0, 1, ..., N-1, \tag{6}$$

при условии

$$u_0 = 0. \tag{7}$$

Последовательность v_n определяем равенствами

$$2v_n - v_{n+1} = 0, \quad n = 0, 1, ..., N-1, \tag{8}$$

$$v_0 = 1. \tag{9}$$

Число c определяем из условия (5). При этом значения u_n, v_n последовательно вычисляем по формулам

$$u_{n+1} = 2u_n - (1 + n^2/N^2), \quad n = 0, 1, ...,$$
$$v_{n+1} = 2^n, \quad n = 0, 1, ...$$

Алгоритм A_2 состоит в том, что u_n определяем как решение системы (6), но вместо условия (7) используем условие $u_N = 0$. Последовательность v_n определяем как решение уравнений (8), но вместо условия (9) используем $v_N = 1$.

а) Проверить, что второй алгоритм устойчив, а первый очень неустойчив.

б) Попытаться провести решение на компьютере, используя поочередно оба метода при $N = 10$, а затем при $N = 100$.

ЧАСТЬ I

ТАБЛИЧНОЕ ЗАДАНИЕ И ИНТЕРПОЛЯЦИЯ ФУНКЦИЙ. КВАДРАТУРЫ

Одним из основных в математике является понятие функции. Функция $y = f(x)$ $(a \leqslant x \leqslant b)$ может быть задана некоторой формулой, например, $y = x^2$, которую можно хранить и использовать в компьютере в виде программы, по которой для каждого фиксированного значения x вычисляется значение $y = x^2$.

Однако, как правило, функция $y = f(x)$ задается приближенно тем или иным конечным набором чисел — некоторой таблицей, обрабатывая которую, можно получить приближенное значение функции при каждом фиксированном x. Этой таблицей может служить, например, конечный набор первых коэффициентов разложения функции в степенной ряд.

Например, для функции

$$e^x, \quad 0 \leqslant x \leqslant 1, \quad e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + ... + \frac{x^n}{n!} + ...,$$

таблицей может служить конечный набор чисел

$$1, \quad \frac{1}{1!}, \quad ..., \quad \frac{1}{n!};$$

n задано.

Чем больше натуральное n, тем точнее можно восстановить функцию по таблице значений n первых коэффициентов ее разложения в степенной ряд, пользуясь расшифровывающей эту таблицу формулой

$$e^x \approx 1 + \frac{x}{1!} + \frac{x^2}{2!} + ... + \frac{x^n}{n!}.$$

Однако обычно таблица значений функции $y = f(x_n)$ получается в результате ее измерения или вычисления в некотором наборе точек $x_0, x_1, ..., x_n \in [a, b]$. Тогда возникает задача восстановления (интерполяции) функции в точках x, не совпадающих с $x_0, x_1, ..., x_n$.

Наиболее употребительны и удобны алгебраическая и тригонометрическая интерполяции. Мы рассмотрим оба этих способа интерполяции. Рассмотрим здесь также задачу вычисления определенных интегралов по таблице функции, поскольку основные способы получения формул (квадратур) для приближенного вычисления определенных интегралов тесно связаны с интерполяционными формулами.

ГЛАВА 1
АЛГЕБРАИЧЕСКАЯ ИНТЕРПОЛЯЦИЯ

Пусть заданы точки $x = x_0$, $x = x_1$, ..., $x = x_n$ и значения $f(x_0)$, $f(x_1), ..., f(x_n)$ функции $f(x)$ в этих точках. Соответствие

x_0	x_1	...	x_n
$f(x_0)$	$f(x_1)$	...	$f(x_n)$

будем называть *таблицей значений функции* $f(x)$ *в узлах* $x_0, x_1, ..., x_n$.

Это название несколько условно, так как значение $f(x_j)$ может записываться бесконечной десятичной дробью (например, $\sqrt{3}$), а для работы на машине все числа должны быть округлены до десятичных (или двоичных) дробей с конечным числом знаков.

Алгебраическим интерполяционным многочленом $P_n(x) = P_n(x, f, x_0, x_1, ..., x_n)$ назовем многочлен

$$P_n(x) = c_0 + c_1 x + ... + c_n x^n$$

степени не выше n, который в узлах $x_0, x_1, ..., x_n$ принимает значения $f(x_0), f(x_1), ..., f(x_n)$.

§ 1. Существование и единственность интерполяционного многочлена

1. Интерполяционный многочлен в форме Лагранжа.

Теорема 1. *Пусть заданы узлы* $x_0, x_1, ..., x_n$, *среди которых нет совпадающих, и значения* $f(x_0), f(x_1), ..., f(x_n)$ *функции в этих узлах. Существует один и только один многочлен* $P_n(x) = P_n(x, f, x_0, x_1, ..., x_n)$ *степени не выше* n, *принимающий в заданных узлах* x_k *заданные значения* $f(x_k)$.

Доказательство. Сначала покажем, что существует не более, чем один интерполяционный многочлен $P_n(x)$, а затем построим его.

Если бы таких многочленов было два, $P_n^{\mathrm{I}}(x)$ и $P_n^{\mathrm{II}}(x)$, то их разностью $R_n(x) = P_n^{\mathrm{I}}(x) - P_n^{\mathrm{II}}(x)$ был бы многочлен степени не выше n, обращающийся в нуль в $n+1$ точках $x_0, x_1, ..., x_n$. Но каждый многочлен, отличный от тождественного нуля, имеет ровно столько корней, считая их кратности, какова его степень. Поэтому $R_n(x) \equiv 0$, т. е. $P_n^{\mathrm{I}}(x) \equiv P_n^{\mathrm{II}}(x)$. Единственность доказана.

Введем теперь вспомогательные многочлены

$$l_k(x) = \frac{(x - x_0)(x - x_1)...(x - x_{k-1})(x - x_{k+1})...(x - x_n)}{(x_k - x_0)(x_k - x_1)...(x_k - x_{k-1})(x_k - x_{k+1})...(x_k - x_n)}.$$

Очевидно, что $l_k(x)$ есть многочлен степени n и что выполнены ра-

венства
$$l_k(x)\big|_{x_j} = \begin{cases} 1, & x_j = x_k, \\ 0, & x_j \neq x_k, \end{cases} \qquad j = 0, 1, ..., n.$$

Многочлен $P_n(x)$, заданный равенством
$$P_n(x) = P_n(x, f, x_0, x_1, ..., x_n) = \\ = f(x_0)l_0(x) + f(x_1)l_1(x) + ... + f(x_n)l_n(x), \quad (1)$$
и есть искомый интерполяционный многочлен.

Действительно, он имеет степень не выше n, так как каждое слагаемое $f(x_j)l_j(x)$ есть многочлен степени не выше n. Кроме того, для него, очевидно, выполнены равенства $P_n(x_j) = f(x_j)$ $(j = 0, 1, ..., n)$. □

Мы не только доказали теорему, но и выписали интерполяционный многочлен $P_n(x)$ в виде формулы (1), которая называется *записью интерполяционного многочлена в форме Лагранжа.*

Употребительны и другие записи (единственного) интерполяционного многочлена $P_n(x, f, x_0, x_1, ..., x_n)$. Особенно часто используют запись в форме Ньютона.

2. Интерполяционный многочлен в форме Ньютона. Разностные отношения. Пусть функция $f(x)$ в точках x_a, x_b, x_c, x_d и т. д. принимает некоторые значения $f(x_a)$, $f(x_b)$, $f(x_c)$, $f(x_d)$ и т. д. Разностное отношение нулевого порядка $f(x_k)$ функции $f(x)$ в точке x_k определим как значение функции в этой точке:
$$f(x_k) = f(x_k), \quad k = a, b, c, d, ... \quad (2)$$

Разностное отношение первого порядка $f(x_k, x_t)$ функции $f(x)$ для (произвольной) пары точек x_k, x_t определим через разностные отношения нулевого порядка:
$$f(x_k, x_t) = \frac{f(x_t) - f(x_k)}{x_t - x_k}.$$

Вообще разностное отношение $f(x_0, x_1, ..., x_n)$ n-го порядка определим через разностные отношения порядка $n-1$, положив
$$f(x_0, x_1, ..., x_n) = \frac{f(x_1, x_2, ..., x_n) - f(x_0, x_1, ..., x_{n-1})}{x_n - x_0}. \quad (3)$$

Интерполяционный многочлен $P_n(x, f, x_0, x_1, ..., x_n)$ можно записать в следующей форме Ньютона:
$$P_n(x, f, x_0, x_1, ..., x_n) = f(x_0) + (x - x_0)f(x_0, x_1) + ... \\ ... + (x - x_0)(x - x_1)...(x - x_{n-1})f(x_0, x_1, ..., x_n). \quad (4)$$

Несколько ниже мы докажем справедливость формулы (4), а пока установим некоторые следствия из нее.

Следствие 1. *Справедливо равенство*
$$P_n(x, f, x_0, x_1, ..., x_{n-1}, x_n) = P_{n-1}(x, f, x_0, x_1, ..., x_{n-1}) + \\ + f(x_0, x_1, ..., x_n)(x - x_0)(x - x_1)...(x - x_{n-1}). \quad (5)$$

Доказательство очевидно.

Следствие 2. *Разностное отношение* $f(x_0, x_1, ..., x_n)$ *порядка* n *равно коэффициенту* c_n *при члене* x^n, *входящем в интерполяционный многочлен*

$$P_n(x, f, x_0, x_1, ..., x_n) = c_n x^n + c_{n-1}x^{n-1} + ... + c_0,$$

т. е. справедливо равенство

$$f(x_0, x_1, ..., x_n) = c_n. \tag{6}$$

Доказательство. Очевидно, что в правую часть выражения (4) член x^n входит с коэффициентом $f(x_0, x_1, ..., x_n)$. □

Следствие 3. *Разностное отношение* $f(x_0, x_1, ..., x_n)$ *обращается в нуль в том и только том случае, если* $f(x_0), f(x_1), ..., f(x_n)$ *суть значения некоторого многочлена* Q_m, *степень которого* m *строго меньше* n.

Доказательство. Если $f(x_0, x_1, ..., x_n) = 0$, то из выражения (4) видно, что интерполяционный многочлен $P_n(x, f, x_0, x_1, ..., x_n)$, принимающий при x_j значения $f(x_j)$ $(j = 0, 1, ..., n)$, есть многочлен степени меньше n, поскольку в силу равенства (6) коэффициент c_n при x^n равен нулю. Обратно: ввиду единственности интерполяционного многочлена степени не выше n многочлен $Q_m(x)$ совпадает с интерполяционным многочленом $P_n(x, f, x_0, ..., x_n) = c_n x^n + c_{n-1}x^{n-1} + ... + c_0$. Ввиду того, что $m < n$, из равенства $Q_m(x) = P_n(x, f, x_0, x_n)$ следует, что $c_n = 0$. В силу равенства (6) тогда $f(x_0, x_1, ..., x_n) = 0$. □

Следствие 4. *Разностное отношение* $f(x_0, x_1, ..., x_n)$ *не изменяется при произвольной перестановке его аргументов* $x_0, x_1, ..., x_n$ *между собой.*

Доказательство. Переставим узлы $x_0, x_1, ..., x_n$ между собой так, что на месте с номером j окажется один из узлов $x_0, x_1, ..., x_n$, который мы обозначим x_j $(j = 0, 1, ..., n)$. Очевидно, что интерполяционный многочлен от нумерации узлов не зависит: $P_n(x, f, x_0, x_1, ..., x_n) \underset{x}{\equiv} \underset{x}{\equiv} P_n(x, f, x'_0, x'_1, ..., x'_n)$. Поэтому наряду с записью (4) справедлива запись

$$P_n(x, f, x_0, x_1, ..., x_n) = f(x'_0) + f(x'_0, x'_1)(x - x'_0) + ... \\ ... + f(x'_0, x'_1, ..., x'_n)(x - x'_0)(x - x'_1)...(x - x'_{n-1}), \tag{7}$$

так что в силу равенства (6)

$$f(x'_0, x'_1, ..., x'_n) = c_n. \tag{8}$$

Сравнивая равенства (6) и (8), убеждаемся в справедливости доказываемого утверждения $f(x_0, x_1, ..., x_n) = f(x'_0, x'_1, ..., x'_n)$. □

Следствие 5. *Справедливо равенство*

$$f(x_0, x_1, ..., x_n) = \frac{f(x_n) - P_{n-1}(x_n, f, x_0, x_1, ..., x_{n-1})}{(x_n - x_0)(x_n - x_1)...(x_n - x_{n-1})}. \tag{9}$$

Доказательство. В равенстве (5) положим $x = x_n$. Тогда левая часть примет значение $f(x_n)$, а формула (9) станет очевидной. □

Теорема 2. *Интерполяционный многочлен* $P_n(x, f, x_0, x_1, ..., x_n)$ *допускает запись в форме Ньютона, т. е. имеет место формула* (4).

Доказательство. Воспользуемся индукцией по n. При $n = 0$ формула (4) справедлива. Допустим, что ее справедливость уже установлена для $n = 1, 2, ..., k$. Покажем, что она имеет место и для $n = k + 1$, т. е. докажем равенство

$$P_{k+1}(x, f, x_0, x_1, ..., x_k, x_{k+1}) = P_k(x, f, x_0, x_1, ..., x_k) + \\ + f(x_0, x_1, ..., x_k, x_{k+1})(x - x_0)(x - x_1)...(x - x_k). \quad (10)$$

Заметим, что в силу предположения индукции о справедливости равенства (4) для $n \leqslant k$ доказательства следствий 1–5, которые мы провели, опираясь на справедливость равенства (4), сохраняют силу при $n \leqslant k$.

Переходя к доказательству равенства (10), сначала покажем, что многочлен $P_{k+1}(x, f, x_0, x_1, ..., x_k, x_{k+1})$ можно записать в форме

$$P_{k+1}(x, f, x_0, x_1, ..., x_k, x_{k+1}) = P_k(x, f, x_0, x_1, ..., x_k) + \\ + \frac{f(x_{k+1}) - P_k(x_{k+1}, f, x_0, x_1, ..., x_k)}{(x_{k+1} - x_0)...(x_{k+1} - x_k)}(x - x_0)...(x - x_k). \quad (11)$$

Очевидно, что в правой части равенства (11) стоит многочлен степени не выше $k + 1$, принимающий в точках x_j значения $f(x_j)$ $(j = 0, 1, ..., k + 1)$. Поэтому выражение в правой части равенства (11) есть интерполяционный многочлен

$$P_{k+1}(x, f, x_0, x_1, ..., x_{k+1}).$$

Сравнивая формулы (10) и (11), видим, что для доказательства справедливости (10) надо установить равенство

$$f(x_0, x_1, ..., x_{k+1}) = \frac{f(x_{k+1}) - P_k(x_{k+1}, f, x_0, x_1, ..., x_k)}{(x_{k+1} - x_0)...(x_{k+1} - x_{k-1})(x_{k+1} - x_k)}. \quad (12)$$

В силу следствия 4 получаем

$$P_k(x, f, x_0, x_1, ..., x_k) = P_k(x, f, x_1, x_2, ..., x_k, x_0) = \\ = P_{k-1}(x, f, x_1, x_2, ..., x_k) + \\ + f(x_1, x_2, ..., x_k, x_0)(x - x_1)(x - x_2)...(x - x_k). \quad (13)$$

Воспользуемся формулой (13) при $x = x_{k+1}$ и придадим правой части равенства (12) следующий вид:

$$\frac{f(x_{k+1}) - P_k(x_{k+1}, f, x_0, x_1, ..., x_k)}{(x_{k+1} - x_0)...(x_{k+1} - x_{k-1})(x_{k+1} - x_k)} = \\ = \frac{1}{x_{k+1} - x_0} \frac{f(x_{k+1}) - P_{k-1}(x_{k+1}, f, x_1, ..., x_k)}{(x_{k+1} - x_1)...(x_{k+1} - x_k)} - \frac{f(x_1, x_2, ..., x_k, x_0)}{x_{k+1} - x_0}. \quad (14)$$

В силу следствия 5 уменьшаемое в правой части равенства (14) совпадает с выражением

$$\frac{1}{x_{k+1}-x_0}\, f(x_1, x_2, \ldots, x_{k+1}).$$

В силу следствия 4 в вычитаемом можно переставить аргументы так, что оно совпадет с $\dfrac{f(x_0, x_1, \ldots, x_k)}{x_{k+1}-x_0}$.

Таким образом, правая часть равенства (14) есть

$$\frac{f(x_1, x_2, \ldots, x_{k+1}) - f(x_0, x_1, \ldots, x_k)}{x_{k+1}-x_0} = f(x_0, x_1, \ldots, x_{k+1}),$$

так что равенство (14) совпадает с доказываемым равенством (12). □

Теорема 3. *Пусть* $x_0 < x_1 < \ldots < x_n$, *функция* $f(x)$ *определена на отрезке* $x_0 \leqslant x \leqslant x_n$ *и имеет на этом отрезке производную порядка* n. *Тогда*

$$n!\, f(x_0, x_1, \ldots, x_n) = f^{(n)}(\xi), \tag{15}$$

где ξ *— некоторая точка отрезка* $[x_0, x_n]$.

Доказательство. Функция

$$\varphi(x) \equiv f(x) - P_n(x, f, x_0, \ldots, x_n) \tag{16}$$

обращается в нуль в $n+1$ точках $x_0, x_1, \ldots, x_n$. По теореме Ролля ее производная обращается в нуль хотя бы в одной точке между каждыми двумя соседними нулями функции $\varphi(x)$. Таким образом, функция $\varphi'(x)$ обращается в нуль не менее, чем в n точках. Аналогично $\varphi''(x)$ обращается в нуль по крайней мере в одной точке между каждыми двумя нулями функции $\varphi'(x)$ и имеет поэтому не менее, чем $n-1$ нулей.

Рассуждая аналогично, убедимся, что $\varphi^{(n)}(x)$ имеет хотя бы один нуль. Обозначим его ξ, так что $\varphi^{(n)}(\xi) = 0$. Продифференцируем тождество (16) ровно n раз и положим после этого $x = \xi$:

$$0 = \varphi^{(n)}(\xi) = f^{(n)}(\xi) - \frac{d^n}{dx^n} P_n(x, f, x_0, x_1, \ldots, x_n)\big|_{x=\xi}. \tag{17}$$

Но

$$\frac{d^n}{dx^n}\left[P_n(x, f, x_0, x_1, \ldots, x_n)\right] = \frac{d^n}{dx^n}\left[P_{n-1}(x, f, x_0, x_1, \ldots, x_{n-1}) + \right.$$
$$\left. + f(x_0, x_1, \ldots, x_n)(x-x_0)(x-x_1)\ldots(x-x_{n-1})\right] \equiv 0 + n!\, f(x_0, x_1, \ldots, x_n).$$

Поэтому из выражения (17) следует равенство (15). □

Теорема 4. *Значения* $f(x_0), f(x_1), ..., f(x_n)$ *выражаются через разностные отношения* $f(x_0), f(x_0, x_1), ..., f(x_0, x_1, ..., x_n)$ *формулами*

$$f(x_i) = f(x_0) + (x_i - x_0)f(x_0, x_1) + (x_i - x_0)(x_i - x_1)f(x_0, x_1, x_2) + ...$$
$$... + (x_i - x_0)(x_i - x_1)...(x_i - x_{n-1})f(x_0, x_1, ..., x_n),$$
$$i = 0, 1, ..., n,$$

т. е. формулами вида

$$f(x_i) = a_{i0}f(x_0) + a_{i1}f(x_0, x_1) + ... + a_{in}f(x_0, ..., x_n), \quad i = 0, 1, ..., n. \tag{18}$$

Для доказательства можно воспользоваться равенствами $f(x_i) = P_n(x, f, x_0, x_1, ..., x_n)|_{x=x_i}$ и записью интерполяционного многочлена в форме (4).

3. Сравнение записей в форме Лагранжа и Ньютона. Для вычисления значения функции $f(x)$ в точке x, не являющейся узлом интерполяции, можно положить $f(x) \approx P_n(x, f, x_0, x_1, ..., x_n)$.

Пусть $P_n(x, f, x_0, x_1, ..., x_n)$ уже найден, но мы решили для уточнения привлечь еще один узел x_{n+1} и значение $f(x_{n+1})$ в нем. Тогда для вычисления $P_{n+1}(x, f, x_0, x_1, ..., x_n, x_{n+1})$ с помощью формулы (1) нужно заново провести всю работу. Для вычисления по формуле Ньютона

$$P_{n+1}(x, f, x_0, x_1, ..., x_n, x_{n+1}) = P_n(x, f, x_0, x_1, ..., x_n) +$$
$$+ f(x_0, x_1, ..., x_{n+1})(x - x_0)(x - x_1)...(x - x_n)$$

нужно вычислить только поправку

$$f(x_0, x_1, ..., x_{n+1})(x - x_0)(x - x_1)...(x - x_n).$$

Кстати, сразу будет видно, насколько она велика.

4. Обусловленность задачи построения интерполяционного многочлена. Пусть узлы интерполяции $x_0, x_1, ..., x_n$ лежат на некотором отрезке $a \leqslant x \leqslant b$. Пусть $f(x_0), f(x_1), ..., f(x_n)$ — заданные числа. Соответствующий интерполяционный многочлен $P_n(x) = P_n(x, f, x_0, x_1, ..., x_n)$ будем для краткости обозначать $P_n(x, f)$.

Придадим значениям $f(x_j)$ некоторые возмущения $\delta f(x_j)$ $(j = 0, 1, ..., n)$, и интерполяционный многочлен $P_n(x, f)$ заменится многочленом $P_n(x, f + \delta f)$. Из записи (1) видно, что $P_n(x, f + \delta f) = P_n(x, f) + P_n(x, \delta f)$, так что возмущение, которое претерпевает интерполяционный многочлен, есть $P_n(x, \delta f)$. Это возмущение при заданных $x_0, x_1, ..., x_n$ зависит только от δf, но не от f. Примем за меру чувствительности интерполяционного многочлена к возмущениям δf задания функции в узлах наименьшее число L_n, при котором для каждого δf выполнено неравенство

$$\max_{a<x<b} |P_n(x, \delta f)| \leqslant L_n \max_j |\delta f(x_j)|.$$

Числа $L_n = L_n(x_0, x_1, ..., x_n, a, b)$ $(n = 0, 1, ...)$ называют *константами Лебега*. Эти числа растут с ростом n. Их поведение при возрастании n существенно зависит от расположения точек x_j $(j = 0, 1, ..., n)$ на отрезке $[a, b]$.

Если, например, $n = 1$, $x_0 = a$, $x_1 = b$, то $L_1 = 1$. Если $x_0 \neq a$, $x_1 \neq$ $\neq b$, то $L_1 \geqslant \dfrac{b-a}{2|x_1 - x_0|}$, т. е. чувствительность интерполяции может быть сколь угодно сильной, если x_1 и x_0 достаточно мало различаются. Читатель легко проверит эти утверждения об L_1.

В случае равномерно расположенных узлов

$$x_j = a + jh, \quad j = 0, 1, ..., n; \quad h = \frac{b-a}{n},$$

можно показать, что

$$2^{n-1} > L_n > 2^{n-3} \frac{1}{\sqrt{n-1}} \cdot \frac{1}{n - 3/2}, \tag{19}$$

т. е. чувствительность результата интерполяции к погрешностям при задании $f(x_j)$ будет резко возрастать с ростом n. Погрешности при задании $f(x_j)$ неизбежны как при получении значений путем измерений, так и в результате округлений.

Пусть теперь $a = -1$, $b = 1$, а узлы заданы формулой

$$x_j = -\cos \frac{(2j+1)\pi}{2(n+1)}, \quad j = 0, 1, ..., n. \tag{20}$$

Можно показать, что при условии (20)

$$L_n \leqslant \frac{2}{\pi} \ln n + 1, \tag{21}$$

т. е. с ростом n константы Лебега, в отличие от (19), растут очень медленно, так что в этом случае вычислительная неустойчивость не является препятствием для использования интерполяционных многочленов высокой степени.

5. О плохой сходимости интерполяции по равноотстоящим узлам. Не следует думать, что для каждой непрерывной функции $f(x)$, $x \in [a, b]$, интерполяционный многочлен $P_n(x, f)$, построенный по значениям $f(x_j)$ в равноотстоящих узлах $x_j = a + jh$, $x_0 = a$, $x_n = b$, с ростом n все меньше уклоняется от функции $f(x)$. Можно показать, например, что для функции $f = \dfrac{1}{x^2 + 0{,}25}$, имеющей производные всех порядков, при $a = -1$ и $b = 1$ уклонение $\max\limits_{-1 \leqslant x \leqslant 1} |f(x) - P_n(x)|$ не стремится к нулю с ростом n.

Задачи

1. Требуется вычислить $f(1,14)$ с помощью линейной, квадратической и кубической интерполяций, используя следующую таблицу:

x	1,08	1,13	1,20	1,27	1,31
$f(x)$	1,302	1,386	1,509	1,217	1,284

Осуществить расчет с помощью интерполяционных многочленов в формах Лагранжа и Ньютона.

2. Пусть $x_j = jh$ $(j = 0, \pm 1, \pm 2, ...)$ — узлы, расположенные с шагом h. Проверить равенство

$$f(x_{k-1}, x_k, x_{k+1}) = \frac{f(x_{k+1}) - 2f(x_k) + f(x_{k-1})}{2!\, h^2}.$$

3. Пусть $a = x_0$, $a < x_1 < b$, $x_2 = b$. Вычислить значение константы Лебега L_2, если $x_1 = (a+b)/2$.

Показать, что при $x_1 \to a$ или $x_1 \to b$ константа Лебега $L_2 = L_2(x_0, x_1, x_2, a, b)$ неограниченно возрастает.

4. Проверить утверждение об интерполяции функции $f(x) = \dfrac{1}{x^2 + 0{,}25}$ из п. 5 экспериментально, выводя графики $f(x)$ и $P_n(x, f)$ на экран дисплея ЭВМ.

§ 2. Классическая кусочно многочленная интерполяция

Высокая чувствительность интерполяционных многочленов к погрешностям при задании таблицы и возможное отсутствие сходимости последовательности $P_n(x, f)$ с ростом n при равноотстоящих узлах заставляет использовать кусочно многочленную интерполяцию.

1. Определение кусочно многочленной интерполяции. Пусть функция $f(x)$, определенная на $[a, b]$, задана в виде $f = (f(x_0), f(x_1), ..., f(x_n))$ $(a = x_0 < x_1 < x_2 < ... < x_n = b)$. Для восстановления функции между узлами $x_0, x_1, ..., x_n$ можно воспользоваться функцией, которая между каждыми двумя соседними узлами является многочленом заданной невысокой степени, например первой, второй, третьей и т. д. Соответствующая интерполяция называется *кусочно линейной*, *кусочно квадратичной* и т. д.

В случае кусочно линейной интерполяции на отрезке $x_k \leqslant x \leqslant x_{k+1}$ для аппроксимации функции $f(x)$ используется линейный интерполяционный многочлен $P_1(x, f, x_k, x_{k+1})$. В случае квадратичной интерполяции на отрезке $x_k \leqslant x \leqslant x_{k+1}$ можно воспользоваться одним из многочленов $P_2(x, f, x_k, x_{k+1}, x_{k+2})$ или $P_2(x, f, x_{k-1}, x_k, x_{k+1})$.

Аналогично строятся кусочно многочленные интерполяции произвольной степени s. Именно, при заданном s на отрезке $[x_k, x_{k+1}]$ используем интерполяционный многочлен $P_s(x, f, x_{k-j}, x_{k-j+1}, ..., x_{k-j+s})$, где j — одно из чисел $0, 1, ..., s-1$. Желательно, чтобы отрезок $[x_k, x_{k+1}]$ был возможно ближе к середине отрезка $[x_{k-j}, x_{k-j+s}]$. Поэтому целесообразно выбирать в качестве j то из чисел $0, 1, ..., s-1$, которое ближе к $s/2$, и вместо $P_s(x, f, x_{k-j}, ..., x_{k-j+s})$ использовать сокращенное обозначение, положив

$$P_s(x, f, x_{k-j}, ..., x_{k-j+s}) = P_s(x, f_{kj}).$$

2. Формула для погрешности интерполяции. Займемся оценкой погрешности

$$R_s(x) \equiv f(x) - P_s(x, f_{kj}), \quad x_k \leqslant x \leqslant x_{k+1}, \tag{1}$$

возникающей при приближенной замене $f(x)$ многочленом $P_s(x, f_{kj})$. Воспользуемся следующей общей теоремой о формуле для погрешности.

Теорема 1. *Пусть $f(t)$ — функция, определенная на некотором отрезке $\alpha \leqslant t \leqslant \beta$ и имеющая непрерывную производную некоторого порядка $s+1$. Пусть $t_0, t_1, ..., t_s$ — произвольный набор попарно различных точек из отрезка $[\alpha, \beta]$, $f(t_0), f(t_1), ..., f(t_s)$ — значения функции $f(t)$ в этих точках, а $P_s(t)$ — интерполяционный многочлен степени не выше s, построенный по этим значениям. Тогда погрешность интерполяции $R_s(t) = f(t) - P_s(t)$ можно представить в виде*

$$R_s(t) = \frac{f^{(s+1)}(\xi)}{(s+1)!}(t-t_0)(t-t_1)...(t-t_s), \tag{2}$$

где $\xi = \xi(t)$ — некоторая точка из интервала $\alpha < t < \beta$.

Доказательство. Фиксируем произвольное $t = \bar{t} \in [\alpha, \beta]$ и докажем формулу (1) при этом $\bar{t}$. Если $\bar{t}$ совпадает с одним из узлов интерполяции, $\bar{t} = t_j$ $(j = 0, 1, ..., s)$, то погрешность $f(\bar{t}) - P_s(\bar{t})$ обращается в нуль. Формула (2) также дает $R_s(\bar{t}) = 0$ при произвольном ξ, так что формула (2) доказана для $t = \bar{t} = t_j$ $(j = 0, 1, ..., s)$.

Докажем теперь формулу (1) для $t = \bar{t}$ в предположении, что фиксированное $\bar{t} \in [\alpha, \beta]$ не совпадает ни с одним из узлов интерполяции. Введем вспомогательную функцию

$$\varphi(t) = f(t) - P_s(t) - k(t-t_0)(t-t_1)...(t-t_s), \tag{3}$$

выбрав число k так, чтобы при $t = \bar{t}$ функция $\varphi(t)$ обращалась в нуль, т. е. положив

$$k = \frac{f(\bar{t}) - P_s(\bar{t})}{(\bar{t}-t_0)(\bar{t}-t_1)...(\bar{t}-t_s)}. \tag{4}$$

Числитель в формуле (4) есть значение $R_s(\bar{t})$ погрешности; поэтому из этой формулы следует, что

$$R_s(\bar{t}) = k(\bar{t}-t_0)(\bar{t}-t_1)...(\bar{t}-t_s). \tag{5}$$

Функция $\varphi(t)$ обращается в нуль в $s+2$ точках $\bar{t}, t_0, t_1, ..., t_s$. Производная $\varphi'(t)$ обращается в нуль не менее, чем в $(s+1)$-й точке, так как по теореме Ролля между каждыми двумя соседними точками, где функция $\varphi(t)$ обращается в нуль, найдется точка, где обращается в нуль ее производная $\varphi'(t)$. Аналогично $\varphi''(t)$ имеет не менее, чем s нулей, $\varphi^{(3)}(t)$ — не менее, чем $s-1$ нулей и т. д., наконец, $(s+1)$-я

производная $\dfrac{d^{s+1}}{dx^{s+1}}\varphi(t)$ обращается в нуль хотя бы в одной точке $\xi \in (\alpha, \beta)$. Заметим, что

$$\frac{d^{s+1}}{dt^{s+1}}t^{s+1} = (s+1)!,$$

а также $(t-t_0)(t-t_1)...(t-t_s) = t^{s+1} + Q_s(t)$, где $Q_s(t)$ — некоторый многочлен степени s. Заметим еще, что

$$\frac{d^{s+1}}{dx^{s+1}}P_s(t) \equiv \frac{d^{s+1}}{dx^{s+1}}Q_s(t) \equiv 0.$$

Возьмем производную порядка $s+1$ от функции $\varphi(t)$, заданной формулой (3). Получим

$$\varphi^{(s+1)}(t) = f^{(s+1)}(t) - k(s+1)!.$$

Отсюда при $t = \xi$ в силу $\varphi^{(s+1)}(\xi) = 0$ следует, что

$$k = \frac{f^{(s+1)}(\xi)}{(s+1)!}.$$

Подставляя найденное k в равенство (5), получаем формулу для $R_s(\bar{t}\,)$, которая в силу произвольности выбора $\bar{t}$ совпадает с формулой (2). □

Теорема 2. *В условиях предыдущей теоремы справедлива оценка*

$$\max_{\alpha \leqslant t \leqslant \beta} |R_s(t)| \leqslant \frac{1}{(s+1)!} \max_{\alpha \leqslant t \leqslant \beta} |f^{(s+1)}(t)|(\beta-\alpha)^{s+1}. \tag{6}$$

Доказательство. Заметим, что для любого значения $t \in [\alpha, \beta]$ каждое из выражений $t-t_0, t-t_1, ..., t-t_s$ не превосходит по модулю число $\beta - \alpha$, а затем воспользуемся формулой (2):

$$|R_s(t)| = \frac{1}{(s+1)!}|f^{(s+1)}(\xi)(t-t_0)(t-t_1)...(t-t_s)| \leqslant$$
$$\leqslant \frac{1}{(s+1)!} \max_{\alpha \leqslant t \leqslant \beta} |f^{(s+1)}(t)|(\beta-\alpha)^{s+1}. \tag{7}$$

Поскольку $t \in [\alpha, \beta]$ в левой части выражения (7) произвольно, то отсюда следует (6). □

Подчеркнем, что оценка (6) доказана нами при произвольном расположении узлов $t_0, t_1, ..., t_s$ на отрезке $[\alpha, \beta]$.

Для конкретного фиксированного расположения узлов она может быть несколько улучшена. Например, в случае линейной интерполяции и расположения узлов t_0, t_1 в концах α, β отрезка $\alpha \leqslant t \leqslant \beta$ получим

$$|R_1(t)| = \left|\frac{f''(\xi)}{2!}(t-\alpha)(t-\beta)\right| \leqslant$$
$$\leqslant \frac{1}{2} \max_{\alpha \leqslant t \leqslant \beta} |f''(t)| \max_{\alpha \leqslant t \leqslant \beta} |(t-\alpha)(t-\beta)| = \frac{1}{8} \max_{\alpha \leqslant t \leqslant \beta} |f''(t)|(\beta-\alpha)^2 \tag{8}$$

и соответственно

$$\max_{\alpha\leqslant t\leqslant\beta}|R_1(t)| \leqslant \frac{1}{8}\max_{\alpha\leqslant t\leqslant\beta}|f''(t)|(\beta-\alpha)^2, \tag{9}$$

в то время как оценка (6) при $s=1$ принимает вид

$$\max_{\alpha\leqslant t\leqslant\beta}|R_1(t)| \leqslant \frac{1}{2}\max_{\alpha\leqslant t\leqslant\beta}|f''(t)|(\beta-\alpha)^2.$$

Воспользуемся теоремами 1, 2 для оценки погрешности (1) кусочно многочленной интерполяции функции $f(x)$ на отрезке $x_k \leqslant x \leqslant \leqslant x_{k+1}$. Положим

$\alpha = x_{k-j},\ \ \beta = x_{k-j+s},\ \ t_0 = \alpha = x_{k-j},\ \ t_1 = x_{k-j+1},\ \ldots,\ t_s = \beta = x_{k-j+s}$. Далее очевидно, что

$$\max_{x_k\leqslant x\leqslant x_{k+1}}|R_s(x,f_{kj})| \leqslant \max_{\alpha\leqslant x\leqslant\beta}|R_s(x,f_{kj})|,$$

а в силу (6) отсюда следует неравенство

$$\max_{x_k\leqslant x\leqslant x_{k+1}}|R_s(x,f_{kj})| \leqslant \\ \leqslant \frac{1}{(s+1)!}\max_{x_{k-j}\leqslant x\leqslant x_{k-j+s}}|f^{(s+1)}(x)|(x_{k-j+s}-x_{k-j})^{s+1}. \tag{10}$$

Если величина $|f^{(s+1)}(x)|$ сильно меняется на отрезке $[a,b]$, то для получения оценки (10), гарантирующей заданную точность, шаг сетки и число $x_{k-j+s}-x_{k-j}$ должны быть меньше там, где $|f^{(s+1)}(x)|$ больше.

В случае равноотстоящих узлов оценка (10) влечет

$$\max_{x_{k-1}\leqslant x\leqslant x_{k+1}}|R_s(x,f_{kj})| \leqslant \frac{s^{s+1}}{(s+1)!}\max_{x_{k-j}\leqslant x\leqslant x_{k-j+s}}|f^{(s+1)}(x)|h^{s+1}, \tag{11}$$

где h — шаг сетки узлов интерполяции.

Отметим особо случай кусочно линейной интерполяции. В этом случае $s=1$, $\alpha=x_k$, $\beta=x_{k+1}$. Опираясь на оценку (9), получаем

$$\max_{x_k\leqslant x\leqslant x_{k+1}}|R_1(x)| \leqslant \frac{1}{8}\max_{x_k\leqslant x\leqslant x_{k+1}}|f''(x)|(x_{k+1}-x_k)^2. \tag{12}$$

Будем считать в дальнейшем, что $x_0, x_1, \ldots, x_n$ — равноотстоящие узлы, так что $x_{k+1}-x_k = h = (b-a)/n$. Из оценки (11) следует, что

$$\max|R_s(x,f_{kj})| \leqslant \mathrm{const}\cdot\max_{x_{k-j}\leqslant x\leqslant x_{k-j+s}}|f^{(s+1)}(x)|h^{s+1}, \tag{13}$$

где постоянная не зависит от шага сетки h.

3. Приближение производных функции, заданной своими значениями в узлах сетки.

Теорема 3. *Пусть $f(x)$ определена на отрезке $[\alpha, \beta]$ и имеет непрерывную производную некоторого порядка $s+1$ на этом отрезке. Пусть $x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s}$ — узлы интерполяции, причем $\alpha = x_{k-j} < \ldots < x_{k-j+s} = \beta$. Для вычисления производных*

$$\frac{d^q f(x)}{dx^q}, \quad q = 1, 2, \ldots, s,$$

на отрезке $x_k \leqslant x \leqslant x_{k+1}$ можно воспользоваться интерполяционным многочленом $P_s(x, f_{kj})$, положив

$$\frac{d^q f(x)}{dx^q} \approx \frac{d^q}{dx^q} P_s(x, f_{kj}), \quad x_k \leqslant x \leqslant x_{k+1}. \tag{14}$$

При этом погрешность удовлетворяет оценке

$$\max_{x_k \leqslant x \leqslant x_{k+1}} \left| \frac{d^q f(x)}{dx^q} - \frac{d^q}{dx^q} P_s(x, f_{kj}) \right| \leqslant$$

$$\leqslant \frac{\max\limits_{x_{k-j} \leqslant x \leqslant x_{k-j+s}} |f^{(s+1)}(x)|}{(s-q+1)!} (x_{k-j+s} - x_{k-j})^{s+1-q}. \tag{15}$$

Доказательство. Функция $\varphi(x) = f(x) - P_s(x, f_{kj})$ обращается в нуль в точках $x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s}$. Поэтому производная функции $\varphi(x)$ обращается в нуль по крайней мере в s точках, поскольку между каждыми двумя нулями функции $\varphi(x)$ по теореме Ролля есть нуль функции $\varphi'(x)$. Аналогично, $d^q\varphi(x)/dx^q$ имеет не менее, чем $s-q+1$ нулей на отрезке $x_{k-j} \leqslant x \leqslant x_{k-j+s}$.

Это значит, что $\dfrac{d^q f(x)}{dx^q}$ и многочлен $\dfrac{d^q}{dx^q} P_s(x, f_{kj})$ степени не выше $s-q$ совпадают в некоторых $s-q+1$ точках, т. е. многочлен $P_s^{(q)}(x, f_{kj})$ является интерполяционным для функции $f^{(q)}(x)$ на отрезке $x_{k-j} \leqslant x \leqslant x_{k-j+s}$ степени не выше $s-q$ по некоторым $s-q+1$ узлам, лежащим на этом отрезке. Функция $f^{(q)}(x)$ имеет производную порядка $s-q+1$:

$$\frac{d^{s-q+1}}{dx^{s-q+1}} f^{(q)}(x) = \frac{d^{s+1}}{dx^{s+1}} f(x).$$

Поэтому можно воспользоваться теоремой 2, приняв $\alpha = x_{k-j}$, $\beta = x_{k-j+s}$, и в силу оценки (6) написать

$$\max_{\alpha \leqslant x \leqslant \beta} |f^{(q)} - P_s^{(q)}(x, f_{kj})| \leqslant$$

$$\leqslant \frac{1}{(s-q+1)!} \max_{\alpha \leqslant x \leqslant \beta} |f^{(s+1)}(x)| (\beta - \alpha)^{s-q+1}, \tag{16}$$

откуда в силу $\alpha \leqslant x_k < x_{k+1} \leqslant \beta$ следует оценка (15). □

4. Оценка неустранимой погрешности при приближении функции по ее значениям в узлах интерполяции и выбор степени кусочно многочленной интерполяции. Пусть функция $f(x)$ определена на отрезке $[0, \pi]$, и пусть заданы ее значения в узлах равномерной сетки $x_k = k\pi/n$ $(k = 0, 1, ..., n)$. По таблице $f(x_0), f(x_1), ...$ $..., f(x_n)$ в принципе нельзя восстановить функцию $f(x)$ точно, потому что различные функции могут совпадать в точках x_k $(k = 0, 1, ...$ $..., n)$, т. е. иметь одинаковые таблицы. Если, например, о функции $f(x)$ кроме ее таблицы известно лишь, что она непрерывна, то ее нельзя восстановить в точке $x \neq x_k$ $(k = 0, 1, ..., n)$ ни с какой гарантированной точностью.

Пусть о функции $f(x)$ известно, что она имеет производную некоторого порядка $s + 1$, причем

$$\max_x |f^{(s+1)}(x)| \leqslant M_s = \text{const}. \tag{17}$$

Укажем две функции из этого класса, $M_s = 1$:

$$f^{\mathrm{I}}(x) = \frac{\sin nx}{n^{s+1}}, \quad f^{\mathrm{II}}(x) = -\frac{\sin nx}{n^{s+1}},$$

для которых таблицы совпадают (обе чисто нулевые):

$$f^{\mathrm{I}}(x_k) = f^{\mathrm{II}}(x_k) = 0, \quad k = 0, 1, ..., n,$$

и которые уклоняются друг от друга на величину порядка h^{s+1}:

$$\max_{0 \leqslant x \leqslant 1} |f^{\mathrm{I}}(x) - f^{\mathrm{II}}(x)| = \max_{0 \leqslant x \leqslant \pi} 2\left|\frac{\sin nx}{n^{s+1}}\right| = 2h^{s+1}.$$

Таким образом, по таблице функции при дополнительном знании лишь оценки (17) в принципе нельзя восстановить функцию всюду на отрезке $0 \leqslant x \leqslant \pi$ с точностью больше $O(h^{s+1})$. Другими словами, погрешность $O(h^{s+1})$ неустранима при восстановлении функции $f(x)$ $(0 \leqslant x \leqslant \pi)$ по ее таблице.

Очевидно, что

$$\max_x \left|\frac{d^q}{dx^q} f^{\mathrm{I}}(x) - \frac{d^q}{dx^q} f^{\mathrm{II}}(x)\right| = 2\,\frac{1}{n^{s-q+1}} = 2h^{s-q+1},$$

так что неустранимая погрешность при восстановлении производной $d^q f(x)/dx^q$ есть $O(h^{s-q+1})$.

Как мы видели, восстановление функции $f(x)$ или ее производной $d^q f(x)/dx^q$ по приближенным формулам (14) имеет погрешность, по порядку совпадающую с неустранимой. Если использовать интерполяцию степени $q < s$, то погрешность будет $O(h^{q+1})$, т. е. произойдет потеря порядка малости погрешности, дополнительная к той $O(h^{s+1})$, которая обусловлена заданием функции ее таблицей и неустранима.

С другой стороны, использование интерполяции степени $q > s$ не может повысить порядок малости (неустранимой) погрешности $O(h^{s+1})$ и ускорить сходимость при $h \to 0$. В этом смысле степень s

кусочно многочленной интерполяции функций, удовлетворяющих условию (17), является оптимальной.

Замечание. Проведенные рассуждения относятся к поведению погрешности при $h \to 0$. При фиксированном h интерполяция некоторой степени $q < s$ может оказаться более точной, чем интерполяция степени s. Кроме того, при задании $f(x_k)$ $(k = 0, 1, ..., n)$ приближенно с некоторым числом десятичных знаков потеря точности при интерполяции за счет приближенного задания $f(x_k)$ $(k = 0, 1, ..., n)$ при увеличении s возрастает (растут константы Лебега (19) из § 1). Поэтому интерполяция высокой степени (выше третьей) употребляется редко.

5. Насыщаемость (гладкостью) кусочно многочленной интерполяции. Пусть $f(x)$ определена на отрезке $[a, b]$, и пусть задана ее таблица $f(x_k)$ в равноотстоящих узлах x_k $(k = 0, 1, ..., n)$; $h = (b - a)/n$. Мы видели, что погрешность кусочно многочленной интерполяции степени s с помощью интерполяционных многочленов $P_s(x, f_{kj})$ на отрезке $x_k \leqslant x \leqslant x_{k+1}$ в случае, если $f^{(s+1)}(x)$ существует и ограничена, имеет порядок $O(h^{s+1})$. Если об $f(x)$ известно лишь, что она имеет ограниченную производную некоторого порядка $q + 1$ $(q < s)$, то неустранимая погрешность при ее восстановлении по таблице есть $O(h^{q+1})$. Можно показать, что при интерполяции с помощью $P_s(x, f_{kj})$ порядок $O(h^{q+1})$ достигается. Если $f(x)$ имеет ограниченную производную порядка $q + 1$ $(q > s)$, то погрешность интерполяции с помощью $P_s(x, f_{kj})$ остается равной $O(h^{s+1})$, т. е. порядок погрешности не реагирует на дополнительную сверх $s + 1$ производных гладкость функции $f(x)$. Это свойство кусочно многочленной интерполяции называют *свойством насыщаемости* (*гладкостью*).

Задачи

1. Каков должен быть шаг h таблицы функции $f(x) = \sin x$ для того, чтобы при кусочно линейной интерполяции погрешность не превосходила 10^{-6}?

2. Каков должен быть шаг h таблицы функции $f(x) = \sin x$, чтобы погрешность при кусочно квадратичной интерполяции не превосходила 10^{-6}?

3. Значения $f(x)$ могут быть измерены в заданной точке с погрешностью $|\delta f| \leqslant 10^{-4}$.

С каким шагом h разумно составить таблицу функции $f(x)$, если восстанавливать функцию по ее таблице с помощью кусочно линейной интерполяции?

4*. Вопрос задачи 3, но для кусочно квадратичной интерполяции.

5. Пусть

$$f'(x) \approx \begin{cases} \dfrac{f(x+h) - f(x)}{h}, & (18) \\ \dfrac{f(x+h) - f(x-h)}{2h}, & (19) \end{cases}$$

и пусть $|f''(x)| \leqslant 1$ и $|f'''(x)| \leqslant 1$.

а) Найти h, чтобы погрешность была меньше 10^{-3}.

б*) Пусть f задана с погрешностью δ. Какова наибольшая точность, достижимая по формулам (18), (19), и как выбрать h?

в*) Показать, что результат, полученный при оптимальном h по (19), неулучшаем по порядку погрешности относительно δ.

§ 3. Кусочно многочленная гладкая интерполяция (сплайны)

Классическая кусочно линейная, кусочно квадратичная и вообще кусочно многочленная интерполяции заданной степени s приводят к интерполирующей функции (интерполянту), которая в узлах интерполяции, вообще говоря, не имеет производной даже первого порядка.

Существует два типа кусочно многочленных интерполянтов — локальные и нелокальные сплайны, обладающие заданным числом производных всюду, включая узлы интерполяции.

1. Локальная интерполяция гладкости s и ее свойства. Пусть заданы узлы x_l интерполяции и значения функции $f(x_l)$ в них. Зададим натуральное (целое положительное) число s, фиксируем натуральное j $(0 \leqslant j \leqslant s-1)$. Каждой точке x_l сопоставим интерполяционный многочлен $P_s(x, f_{lj})$, построенный по значениям $f(x_{l-j})$, $f(x_{l-j+1})$, ..., $f(x_{l-j+s})$ в узлах $x_{l-j}, x_{l-j+1}, ..., x_{l-j+s}$. Кусочно многочленный локальный сплайн $\varphi(x, s)$, имеющий непрерывные производные порядка s, определим равенствами

$$\varphi(x, s) = Q_{2s+1}(x, k), \quad x \in [x_k, x_{k+1}], \quad k = 0, \pm 1, ..., \tag{1}$$

где $Q_{2s+1}(x, k)$ — многочлен степени не выше $2s+1$, определяемый равенствами

$$\left.\frac{d^m Q_{2s+1}(x, k)}{dx^m}\right|_{x=x_k} = \left.\frac{d^m P_s(x, f_{kj})}{dx^m}\right|_{x=x_k}, \quad m = 0, 1, 2, ..., s, \tag{2}$$

$$\left.\frac{d^m Q_{2s+1}(x, k)}{dx^m}\right|_{x=x_{k+1}} = \left.\frac{d^m P_s(x, f_{k+1,j})}{dx^m}\right|_{x=x_{k+1}}, \quad m = 0, 1, 2, ..., s. \tag{3}$$

Теорема 1. *Существует один и только один многочлен степени не выше $2s+1$, удовлетворяющий* (2), (3).

Доказательство. Система $2s+2$ линейных уравнений относительно коэффициентов многочлена

$$Q_{2s+1}(x, k) = c_{0k} + c_{1k}x + ... + c_{2s+1,k}x^{2s+1},$$

получающегося из системы (2), (3) при замене правых частей нулями, имеет только тривиальное решение

$$c_{ik} \underset{i}{\equiv} 0.$$

Действительно, в противном случае существовал бы многочлен Q_{2s+1} степени не выше $2s+1$, который в силу (2), (3) имел бы корни $x = x_k$, $x = x_{k+1}$ кратности $s+1$ каждый, т. е. имел бы $2s+2$ корней, считая их кратности. В силу этого противоречия определитель системы (2), (3) отличен от нуля, и она имеет одно и только одно решение при любых правых частях. □

Теорема 2. *Пусть $f(x)$ есть многочлен степени не выше s. Тогда интерполянт $\varphi(x,s)$ совпадает с этим многочленом.*

Доказательство. Докажем тождество $\varphi(x,s) = f(x)$ на отрезке $x_k \leqslant x \leqslant x_{k+1}$ между соседними узлами, т. е. докажем, что $Q_{2s+1}(x,k) = f(x)$. В силу единственности интерполяционного многочлена $P_s(x, f_{kj}) \equiv P_s(x, f_{k+1,j}) \equiv f(x)$. Теперь видно, что многочлен $Q_{2s+1}(x,k) \equiv f(x)$ дает решение системы (2), (3). □

Теорема 3. *Кусочно многочленная интерполирующая функция $\varphi(x,s)$, определенная равенством* (1), *в узлах интерполяции x_l совпадает с заданными в них значениями $f(x_l)$ $(l = 0, \pm1, \ldots)$. Кроме того, $\varphi(x,s)$ имеет всюду в области своего определения непрерывную производную порядка s.*

Доказательство. В произвольном узле x_l функции $Q_{2s+1}(x, l-1)$, $Q_{2s+1}(x,l)$ в силу равенств (2), (3) имеют производные порядков $m = 0, 1, \ldots, s$, совпадающие с соответствующими производными одного и того же интерполяционного многочлена $P_s(x, f_{lj})$. В силу (1) это доказывает теорему. □

Запишем $Q_{2s+1}(x,k)$ в виде

$$Q_{2s+1}(x,k) = P_s(x, f_{kj}) + R_{2s+1}(x,k), \tag{4}$$

обозначив через $R_{2s+1}(x,k)$ поправку к классическому интерполяционному многочлену $P_s(x, f_{kj})$.

Теорема 4. *Поправку $R_{2s+1}(x,k)$ можно записать в виде*

$$R_{2s+1}(x,k) = \\ = (x_{k+1} - x_k)^{s+1} f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1})\, q_{2s+1}\Big(\frac{x - x_k}{x_{k+1} - x_k}, k\Big), \tag{5}$$

где $f(x_{k-j}, x_{k-j+1}, \ldots, x_{k-j+s+1})$ — разностное отношение порядка $s+1$ и

$$q_{2s+1}(X,k) = \Big(\frac{x_{k+s-j+1} - x_{k-j}}{x_{k+1} - x_k}\Big) \times \\ \times \sum_{T=0}^{s} \Big\{ \Big[\prod_{i=1}^{s} \Big(X - \frac{x_{k-j+i} - x_k}{x_{k+1} - x_k} \Big) \Big]^{(r)}_{X=1} \Big\} l_r(X), \quad X = \frac{x - x_k}{x_{k+1} - x_k}, \tag{6}$$

$$l_r(X) = \frac{X^{s+1}(X-1)^r}{r!\, s!} \sum_{m=0}^{s-r} (-1)^m \frac{(s+m)!}{m!} (X-1)^m. \tag{7}$$

Выражение $[...]^{(r)}_{X=1}$ означает производную по X порядка r, вычисленную при $X = 1$.

Замечание 1. Можно считать, что формула локальной интерполяции гладкости s получена путем замены в интерполяционном многочлене $P_{s+1}(x, f_{kj})$, записанном в форме Ньютона

$$P_{s+1}(x, f_{kj}) = P_s(x, f_{kj}) + f(x_{k-j}, x_{k-j+1}, ..., x_{k-j+s+1})\,\varphi_{s+1}(x, k),$$

$$\varphi_{s+1}(x, k) = (x - x_{k-j})(x - x_{k-j+1})...(x - x_{k-j+s}),$$

многочлена $\varphi_{s+1}(x, k)$ многочленом

$$(x_{k+1} - x_k)^{s+1} q_{2s+1}\left(\frac{x - x_k}{x_{k+1} - x_k}, k\right).$$

Для иллюстрации теоремы 4 заметим, что в силу формул (4)–(7) в случае $s = 0$, $j = 0$ локальный сплайн $\varphi(x, s)$ осуществляет кусочно линейную интерполяцию. В наиболее интересном для приложений случае $s = 2, j = 1$ и $x_{k+j} - x_k = h = \text{const}$

$$R_5(x, k) = P_2(x, f_{k,1}) + \frac{h^3}{2!}\frac{f(x_{k+2}) - 3f(x_{k+1}) + 3f(x_k) - f(x_{k-1})}{h^3}\times$$
$$\times\left(\frac{x - x_k}{h}\right)^3 \frac{x - x_{k+1}}{h}\left(3 - \frac{2(x - x_k)}{h}\right). \quad (8)$$

Доказательство теоремы 4 приведем в конце параграфа.

Теорема 5. *Пусть функция $f(x)$ определена всюду на отрезке $x \in [x_{k-j}, x_{k-j+s+1}]$ и имеет на этом отрезке ограниченную производную порядка $s + 1$. Тогда на отрезке $x \in [x_k, x_{k+1}]$ справедливы приближенные равенства*

$$f^{(m)}(x) \approx \frac{d^m \varphi(x, s)}{dx^m}, \quad m = 0, 1, ..., s, \quad (9)$$

с оценками погрешности

$$\left| f^{(m)}(x) - \frac{d^m \varphi(x, s)}{dx^m} \right| \leqslant$$
$$\leqslant \text{const}\,\frac{(x_{k+s-j+1} - x_{k-j})^{s+1}}{(x_{k+1} - x_k)^m} \max_{x_{k-j} \leqslant x \leqslant x_{k-j+s+1}} |f^{(s+1)}(x)|, \quad (10)$$
$$m = 0, 1, ..., s.$$

Доказательство. В силу равенства (1) и формулы (4)

$$\left| f^{(m)}(x) - \frac{d^m \varphi(x, s)}{dx^m} \right| \leqslant |f^{(m)}(x) - P_s^{(m)}(x, f_{kj})| + |R^{(m)}_{2s+1}(x, k)|. \quad (11)$$

Но в теореме 3 из § 2 установлена оценка

$$|f^{(m)}(x) - P_s^{(m)}(x, f_{kj})| \leqslant \text{const}_1 \cdot h^{s+1-m}. \quad (12)$$

Далее, учитывая, что $\dfrac{d^m}{dx^m} = \dfrac{1}{(x_{k+1}-x_k)^m}\dfrac{d^m}{dX^m}$, $X = \dfrac{x-x_k}{x_{k+1}-x_k}$, получаем

$$\left|\frac{d^m R_{2s+1}(x,k)}{dx^m}\right| =$$
$$= (x_{k+1}-x_k)^{s+1-m}\,|f(x_{k-j},x_{k-j+1},\dots,x_{k-j+s+1})|\times$$
$$\times\left|\frac{x_{k+s-j+1}-x_{k-j}}{x_{k+1}-x_k}\right|\left|\sum_{r=0}^{s}\left\{\left[\prod_{i=1}^{s}\left(X-\frac{x_{k-j+i}-x_k}{x_{k+1}-x_k}\right)\right]^{(r)}_{X=1}\right\} l_r^{(m)}(X)\right|. \quad (13)$$

В силу теоремы 3 из § 1

$$(s+1)!f(x_{k-j},x_{k-j+1},\dots,x_{k-j+s+1}) = f^{(s+1)}(\xi), \qquad \xi \in [x_{k-j}, x_{k-j+s+1}]. \quad (14)$$

Тогда из (13) очевидно, что

$$\left|\frac{d^m R_{2s+1}(x,k)}{dx^m}\right| \leqslant$$
$$\leqslant \mathrm{const}_2\cdot|x_{k+1}-x_k|^{s+1-m}\max_{x_{k-j}\leqslant x\leqslant x_{k-j+s-1}}\left|\frac{d^{s+1}f(x)}{dx^{s+1}}\right|, \quad (15)$$

где const_2 зависит только от величины отношения

$$\frac{x_{k-j+s+1}-x_{k-j}}{x_{k+1}-x_k} \quad (16)$$

и остается ограниченной, если величина (16) в процессе измельчения сетки остается ограниченной.

Из выражений (11), (12), (15) следует оценка (10). □

Сделаем два замечания о неулучшаемости построенных нами локальных сплайнов (1) в некотором смысле.

З а м е ч а н и е 2. Оценки (10) в случае постоянного шага $x_{k+1} - x_k = (b-a)/n = h$ гарантируют сходимость $\varphi^{(m)}(x,s)$ к $f^{(m)}(x)$ с порядком h^{s+1-m} $(m = 0,1,\dots,s)$. Предполагая, что $f(x)$ имеет ограниченные производные только до порядка $s+1$, более точно восстановить функцию $f(x)$ и ее производные по значениям $f(x_l)$ $(l=0,\pm1,\dots)$ нельзя, поскольку эти значения не содержат достаточной для этого информации, как мы показали в п. 4 § 2.

З а м е ч а н и е 3. Для вычисления $\varphi(x,s)$ при каждом x используется не более $s+2$ узлов интерполяции. Эта характеристика локальности формулы (1) неулучшаема в следующем смысле.

Если потребовать, чтобы использовалось не более $s+1$ точек, то для сходимости $\varphi^{(m)}(x,s)$ к $f^{(m)}(x)$ с порядком h^{s+1-m} пришлось бы отказаться от условия непрерывности даже первых производных от $\varphi(x,s)$ и ограничиться классической кусочно многочленной интерполяцией степени не выше s.

В случае если задано конечное число $n+1$ узлов $x_0, x_1, ..., x_n$, лежащих на отрезке $a \leqslant x \leqslant b$, $a = x_0 < x_1 < ... < x_n = b$, формулу (1) вблизи концов отрезка надо изменить, определив ннтерполянт $\varphi(x) = \varphi(x, s, a, b)$ равенствами

$$\varphi(x,s,a,b) = \begin{cases} P_s(x, f_{jj}), & \text{если} \quad x_0 \leqslant x \leqslant x_j, \\ Q_{2s+1}(x,k), & \text{если} \quad x \in [x_k, x_{k+1}], \\ & \text{причем} \quad k = j, j+1, ..., n+j-s-1, \\ P_s(x, f_{n+j-s,j}), & \text{если} \quad x_{n+j-s} \leqslant x \leqslant x_n. \end{cases}$$

Напомним, что $P_n(x, f_{kj})$ введено в конце п. 1 § 2.

Интерполянт $\varphi(x, s, a, b)$, определенный на $[a, b]$, имеет непрерывные производные до порядка s. Для него имеют смысл утверждения, аналогичные теоремам 1–5.

Локальная гладкая интерполяция введена автором в 1952 г. для функций на многомерной прямоугольной сетке с постоянным шагом; приведенные здесь формулы получены автором в 1974 г. Результаты для многомерного случая и библиографию см. в [16, с. 303–309].

2. Нелокальная гладкая кусочно многочленная интерполяция. Пусть заданы узлы $x_0, x_1, ..., x_n$, $a = x_0 < x_1 < ... < x_n = b$, и значения функции $f(x_i)$ $(i = 0, 1, ..., n)$ в этих узлах. Поставим задачу: найти на каждом отрезке $x_k \leqslant x \leqslant x_{k+1}$ кубический многочлен $P_3(x, k)$ так, чтобы возникающая при этом на отрезке $a \leqslant x \leqslant b$ кусочно многочленная функция совпадала с заданной функцией в узлах и имела непрерывные производные до порядка $s = 2$. Эта функция зависит от двух произвольных постоянных и называется *кубическим сплайном Шонберга*. Сплайны минимальной для заданного s степени построены Шонбергом не только при $s = 2$, но и при всех натуральных s.

Пусть $f(x)$ имеет на $[a, b]$ ограниченную производную некоторого порядка $s+1$. Тогда непрерывный с производными порядка s сплайн при подходящих постоянных сохраняет неулучшаемые приближающие свойства классической, а также локальной гладкой кусочно многочленной интерполяций в том смысле, что для него в случае, если существует $f^{(s+1)}(x)$, выполнены оценки типа (10). Однако сплайны Шонберга теряют свойство локальности, присущее как классической кусочно многочленной интерполяции, так и локальной гладкой интерполяции: коэффициенты многочлена, задающего интерполянт на каком-либо отрезке $x_k \leqslant x \leqslant x_{k+1}$, зависят от значений $f(x_0), f(x_1), ...$ $..., f(x_n)$ функции во всех узлах $x_0, x_1, ..., x_n$, число которых n растет в случае измельчения сетки.

Отметим еще, что (нежелательное) свойство насыщаемости гладкостью, неизбежно присущее локальным классической и гладкой интерполяциям, несмотря на потерю локальности, остается также и у

нелокальных сплайнов: для функций $f(x)$, имеющих более $s+1$ производных, нелокальные сплайны гладкости s при постоянном шаге $h=(b-a)/n$ сетки имеют погрешность $O(h^{s+1})$ того же порядка малости по отношению к $h=1/n$, что и для функций $f(x)$, имеющих только $(s+1)$-ю производную.

Нелокальные сплайны Шонберга при заданном числе непрерывных производных реализуются многочленами наименьшей возможной степени n (например, $m=3$, если $s=2$ — кубические сплайны Шонберга).

Подробнее о сплайнах Шонберга и других сплайнах см. [1, 2, 9, 12, 16, 22] и имеющуюся там библиографию.

3. Доказательство теоремы 4. Коэффициенты многочлена

$$Q_{2s+1}(x,k)=c_0+c_1x+\ldots+c_{2s+1}x^{2s+1} \tag{17}$$

определяются из системы линейных уравнений (2), (3). Правые части уравнений, составляющих систему (2), имеют вид

$$a_0^{(m)}f_{k-j}+a_1^{(m)}f_{k-j+1}+\ldots+a_s^{(m)}f_{k-j+s},$$

а составляющих систему (3) — вид

$$b_0^{(m)}f_{k-j+1}+b_1^{(m)}f_{k-j+2}+\ldots+b_s^{(m)}f_{k-j+s+1},\quad m=0,1,\ldots,s,$$

где $a_i^{(m)}$, $b_i^{(m)}$ $(i=0,1,\ldots,s)$ — некоторые числа, не зависящие от $f_{k-j},\ f_{k-j+1},\ldots,f_{k-j+s+1}$.

Поэтому решение $c_0,c_1,\ldots,c_{2s+1}$ системы (2), (3) состоит из чисел, которые определяются заданными $f_{k-j},f_{k-j+1},\ldots,f_{k-j+s+1}$ по формулам вида

$$c_r=\alpha_0^{(r)}f_{k-j}+\alpha_1^{(r)}f_{k-j+1}+\ldots+\alpha_{s+1}^{(r)}f_{k-j+s+1},\quad r=0,1,\ldots,2s+1, \tag{18}$$

$\alpha_l^{(r)}$ — некоторые числа, не зависящие от $f_{k-j},f_{k-j+1},\ldots,f_{k-j+s+1}$. Подставив выражения (18) в (17) и собрав все слагаемые, содержащие f_{k-j+l} $(l=0,1,\ldots,s+1)$, получим запись многочлена (17) в виде

$$Q_{2s+1}(x,k)=\sum_{l=0}^{s+1}f_{k-j+l}\,p(x,l), \tag{19}$$

где $p(x,l)$ $(l=0,1,\ldots,s+1)$ — некоторые многочлены от x. Заменим в (19) числа f_{k-j+l} по формулам (18) из § 1. Многочлен $Q_{2s+1}(x,k)$ примет вид

$$Q_{2s+1}(x,k)=\sum_{l=0}^{s}f(x_{k-j},x_{k-j+1},\ldots,x_{k-j+l})\,q_l(x,k)+$$
$$+(x_{k+1}-x_k)^{(s+1)}f(x_{k-j+1},x_{k-j+2},\ldots,x_{k-j+s+1})\,\widetilde{q}_{2s+1}(x,k), \tag{20}$$

где $q_l(x,k)$ $(l=0,1,\ldots,s)$ и $\widetilde{q}_{2s+1}(x,k)$ — некоторые многочлены, не зависящие от $f(x_{k-j}),f(x_{k-j+1}),\ldots,f(x_{k-j+s+1})$. Воспользуемся этой

независимостью для отыскания многочленов $q_l(x,k)$. Зададим f_{k-j}, $f_{k-j+1}, ..., f_{k-j+s}$ произвольно, а $f_{k-j+s+1}$ зададим равенством

$$f_{k-j+s+1} = P_s(x, f_{kj})\big|_{x=x_{k-j+s+1}}.$$

В силу теоремы 2 в этом случае

$$Q_{2s+1}(x,k) = P_s(x, f_{kj}),$$

а в силу следствия 3 из формулы Ньютона (см. § 1) разностное отношение $f(x_{k-j}, x_{k-j+1}, ..., x_{k-j+s+1})$ порядка $s+1$ обращается в нуль. Таким образом, формула (20) примет вид

$$P_s(x, f_{kj}) = \sum_{l=0}^{s} f(x_{k-j}, x_{k-j+1}, ..., x_{k-j+l})\, q_l(x,k). \tag{21}$$

В силу произвольности значений $f_{k-j}, f_{k-j+1}, ..., f_{k-j+s}$ отсюда следует, что $q_1(x,k) = (x - x_{k-j})(x - x_{k-j+1})...(x - x_{k-j+l})$ $(l = 0, 1, ...$ $..., s)$. Из (21) следует также, что равенству (20) можно придать вид

$$Q_{2s+1}(x,k) = P_s(x, f_{kj}) + \\ + (x_{k+1} - x_k)^{s+1} f(x_{k-j}, x_{k-j+1}, ..., x_{k-j+s+1})\, \widetilde{q}_{2s+1}(x,k). \tag{22}$$

Для вычисления многочлена $\widetilde{q}_{2s+1}(x,k)$ зададим

$$f_{k-j} = f_{k-j+1} = ... = f_{k-j+s} = 0, \quad f_{k-j+s+1} = 1. \tag{23}$$

При условиях (23) формула (22) примет вид

$$Q_{2s+1}(x,k) = (x_{k+1} - x_k)^{s+1} f(x_{k-j}, x_{k-j+1}, ..., x_{k-j+s+1})\, \widetilde{q}_{2s+1}(x,k). \tag{24}$$

Заметим, что при условиях (23) интерполяционный многочлен $P_{s+1}(x, f, x_{k-j}, x_{k-j+1}, ..., x_{k-j+s+1})$ в силу справедливости его записи в форме Лагранжа совпадает с многочленом

$$\prod_{l=0}^{s} \frac{x - x_{k-j+l}}{x_{k-j+s+1} - x_{k-j+l}}, \tag{25}$$

а в силу записи в форме Ньютона (см. (4) из § 1) имеет вид

$$f(x_{k-j}, x_{k-j+1}, ..., x_{k-j+s+1}) \prod_{l=0}^{s} (x - x_{k-j+l}). \tag{26}$$

Приравнивая (25) и (26) друг другу, получаем, что при условиях (23)

$$f(x_{k-j}, x_{k-j+1}, ..., x_{k-j+s+1}) = \left[\prod_{l=0}^{s} (x_{k-j+s+1} - x_{k-j+l})\right]^{-1}. \tag{27}$$

Таким образом, из (24) получаем

$$\widetilde{q}_{2s+1}(x,k) = (x_{k+1} - x_k)^{-s-1} Q_{2s+1}(x,k) \prod_{l=0}^{s} (x_{k-j+s+1} - x_{k-j+l}), \tag{28}$$

где выражения в правой части вычислены при условиях (23).

Преобразуем (28) к требуемому виду (6), (7). Для этого найдем явную формулу для $Q_{2s+1}(x,k)$ при условиях (23). Заметим, что благодаря условию (2) значение $x = x_k$ является корнем кратности $s+1$ многочлена $Q_{2s+1}(x,k)$, так что $Q(x) = Q_{2s+1}(x)/(x-x_k)^{s+1}$ есть многочлен степени s.

Разложим $Q(x)$ по степеням $x - x_{k+1}$:

$$Q_{2s+1}(x,k) = (x-x_k)^{s+1}\frac{Q_{2s+1}(x,k)}{(x-x_k)^{s+1}} = \\ = (x-x_k)^{s+1}\sum_{r=0}^{s}\frac{1}{r!}\left[\frac{Q_{2s+1}(x,k)}{(x-x_k)^{s+1}}\right]^{(r)}_{x=x_{k+1}}(x-x_{k+1})^r. \quad (29)$$

Используем равенство (3) и напишем

$$\frac{d^r}{dx^r}\left[\frac{Q_{2s+1}(x,k)}{(x-x_k)^{s+1}}\right]\Big|_{x=x_{k+1}} = \\ = \sum_{l=0}^{r}C_r^l\left(\frac{d^{r-l}}{dx^{r-l}}P_s(x,f,x_{k-j+1},\ldots,x_{k-j+s+1})\right)\Big|_{x=x_{k+1}}\times \\ \times[(x-x_k)^{-s-1}]^{(l)}_{x=x_{k+1}}. \quad (30)$$

Но при условии (23) в силу (1)

$$P_s(x,f,x_{k-j+1},\ldots,x_{k-j+s+1}) = \prod_{i=1}^{s}\frac{x-x_{k-j+i}}{x_{k-j+s+1}-x_{k-j+i}}. \quad (31)$$

Подставляя (31) в (30), получаем

$$\frac{d^r}{dx^r}\left[\frac{Q_{2s+1}(x,k)}{(x-x_k)^{s+1}}\right]\Big|_{x=x_{k+1}} = \left[\prod_{i=1}^{s}(x_{k-j+s+1}-x_{k-j+i})\right]^{-1}\times \\ \times\sum_{l=0}^{r}C_r^l[(x-x_k)^{-s-1}]^{(l)}_{x=x_{k+1}}\left[\prod_{i=1}^{s}(x-x_{k-j+i})\right]^{r-l}_{x=x_{k+1}}. \quad (32)$$

Подставляя (32) в (29), а (29) в (28), получаем явное выражение для $\widetilde{q}_{2s+1}(x,k)$ в виде двойной суммы по l и по r, которое зависит от x только как сложная функция $q_{2s+1}(X,k)$ от $X = (x-x_k)/(x_{k+1}-x_k)$:

$$\widetilde{q}_{2s+1}(x,k) = q_{2s+1}(X,k).$$

Изменяя порядок суммирования в полученной формуле для $q_{2s+1}(X,k)$, получаем запись в форме (6), (7). □

Задачи

1. Пусть $x_{k+1} - x_k = h = \text{const}$.

Выпишите многочлены $l_t(x,s)$, с помощью которых локальный сплайн $\widetilde{\psi}(x,s)$ заданной гладкости s, определяемый при $x_k \leqslant x \leqslant x_{k+1}$ равенством (4), можно записать в форме

$$\widetilde{\psi}(x,s) = \sum_t f(x_t)\, l_t(x,s).$$

Рассмотреть случаи: $s=0,\ j=0$; $s=1,\ j=0$; $s=2,\ j=2$.

2. Найти точные значения констант Лебега

$$L_s = \max_{\substack{|f_m|=1,\\ k-j\leqslant m\leqslant k-j+s+1}} \max_{x_k\leqslant x\leqslant x_{k+1}} |Q_{2s+1}(x)|$$

для локальных гладких сплайнов при $s=2,\ j=1$.

3. Доказать, что в случае $x_{t+1} - x_t = h = \text{const}$

$$\max_{x_k\leqslant x\leqslant x_{k+1}} \left| \frac{d^m Q_{2s+1}(x)}{dx^{(m)}} \right| \leqslant$$
$$\leqslant c \max_{0\leqslant j\leqslant s-m+1} |f(x_{k-j+i}, x_{k-j+i+1}, \ldots, x_{k-j+i+m})|, \quad m=0,1,\ldots,s,$$

где $c = c(s,j,m)$ не зависят от h.

Вычислить c в случаях: $s=0,\ j=0$; $s=2,\ j=1$.

§ 4. Интерполяция функций двух переменных

Пусть функция двух переменных $f(x,y)$ задана в узлах некоторой правильной или нерегулярной сетки. Как восстановить ее приближенно в точках (x,y), не принадлежащих множеству узлов?

1. Случай прямоугольной сетки. Пусть сетка образована пересечением прямых $x = x_k$ $(k = 0, \pm1, \ldots)$ и прямых $y = y_l$ $(l = 0, \pm1, \ldots)$. Считаем, что $x_{k+1} > x_k$, $y_{l+1} > y_l$ при любых целых k, l. Значение функции в узле (x_k, y_l) обозначим через f_{kl}. Для вычисления функции в точке $(\overline{x}, \overline{y})$ можно воспользоваться аппаратом кусочно многочленной интерполяции заданной степени s для функций одного переменного.

Для этого сначала осуществляется кусочно многочленная интерполяция заданной степени по x на каждой из прямых $y = y_l$. Затем при каждом интересующем нас значении $x = \overline{x}$ осуществляется кусочно многочленная интерполяция (той же или другой степени) по y вдоль прямой $x = \overline{x}$ по значениям функции $f(x,y)$ в точках $(\overline{x}, y_l)$, полученным на первом шаге процесса. Например, в случае кусочно линейной интерполяции по обоим аргументам этот процесс приводит для интерполяции в прямоугольнике $x_k \leqslant x \leqslant x_{k+1}$, $y_l \leqslant y \leqslant y_{l+1}$ к интерполяционному многочлену

$$P(x,y) = f_{kl}\frac{(x-x_{k+1})(y-y_{l+1})}{(x_k-x_{k+1})(y_l-y_{l+1})} + f_{k+1,l}\frac{(x-x_k)(y-y_{l+1})}{(x_{k+1}-x_k)(y_l-y_{l+1})} +$$

$$+ f_{k+1,l+1}\frac{(x-x_k)(y-y_l)}{(x_{k+1}-x_k)(y_{l+1}-y_l)} + f_{k,l+1}\frac{(x-x_{k+1})(y-y_l)}{(x-x_{k+1})(y_{l+1}-y_l)}.$$

2. Треугольные сетки. Пусть функция $f(x,y)$ определена в некоторой криволинейной области D, причем известно, что в “горловине” функция быстро изменяется.

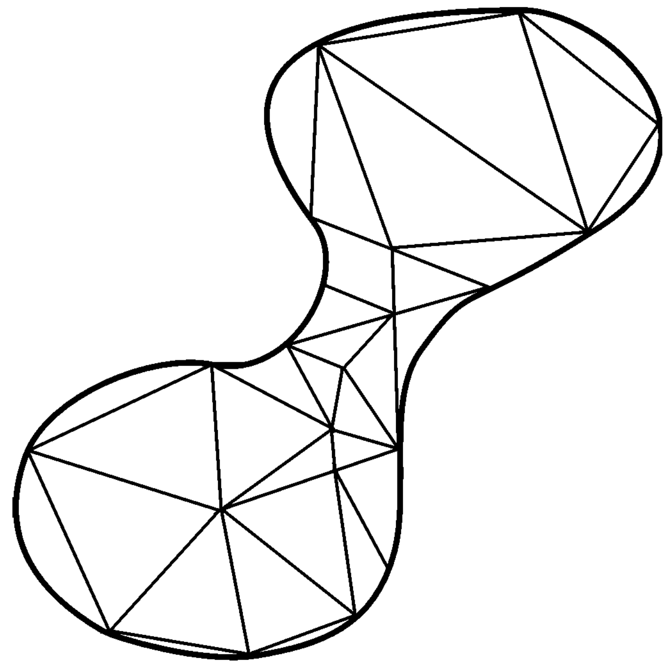


Рис. 2

В таком случае прямоугольная сетка неудобна для табличного задания функции, так как эта сетка не связана с формой границы области D и, кроме того, ее нельзя сгустить только в горловине. Можно воспользоваться треугольной сеткой (рис. 2), свободной от этих недостатков, и принять за узлы интерполяции вершины треугольников.

Если $f_\alpha, f_\beta, f_\gamma$ — значения $f(x,y)$ в вершинах α, β, γ треугольника, то можно осуществить приближенное вычисление функции внутри треугольника с помощью линейной функции

$$f(x,y) \approx P(x,y) = ax + by + c,$$

подобрав a, b, c из условий

$$ax_\alpha + by_\alpha + c = f_\alpha,$$
$$ax_\beta + by_\beta + c = f_\beta,$$
$$ax_\gamma + by_\gamma + c = f_\gamma,$$

где (x_α, y_α), (x_β, y_β), (x_γ, y_γ) — соответственно координаты вершин α, β, γ. Погрешность интерполяции для функции $f(x,y)$ с непрерывными вторыми производными будет $O(h^2)$, где h — длина наибольшей стороны треугольника.

Существуют способы построения кусочно многочленной интерполяции более высокого порядка, а также кусочно многочленной гладкой интерполяции. Аналогичные конструкции существуют и для интерполяции функций многих переменных.

Подчеркнем, что объем таблиц, обеспечивающих возможность восстановления функции данной гладкости с заданной точностью, быстро возрастает с ростом числа аргументов, а алгоритмы усложняются.

Задачи

1. Пусть функция $z = f(x,y)$ имеет вторые производные, ограниченные по модулю единицей.

Как экономно выбрать точки сетки на плоскости, чтобы по значениям функции $z = f(x,y)$ в выбранных точках можно было восстановить функцию в любой точке квадрата $|x| \leqslant 1$, $|y| \leqslant 1$ с погрешностью, не превосходящей $\varepsilon = 10^{-3}$?

2. Решить предыдущую задачу при дополнительном предположении, что окружности $x^2 + y^2 = r^2$ при каждом r являются линиями уровня функции $z = f(x, y)$.

3*. Решить задачу 1 в дополнительном предположении, что функция $z = f(x, y)$ есть функция вида $z = \varphi(x)\psi(y)$, где $\varphi(x)$, $\psi(y)$ — некоторые функции одного аргумента, не обращающиеся в нуль.

ГЛАВА 2
ТРИГОНОМЕТРИЧЕСКАЯ ИНТЕРПОЛЯЦИЯ

Наряду с алгебраической интерполяцией, изложенной в гл. 1, широко употребляется интерполяция с помощью тригонометрических многочленов вида

$$Q\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x\right) = \sum_{k=0}^{n} a_k \cos\frac{2\pi k}{L}x + \sum_{k=1}^{n} b_k \sin\frac{2\pi k}{L}x. \qquad (1)$$

Здесь n — натуральные числа, L — положительное число, a_k, b_k — числовые коэффициенты.

Тригонометрический многочлен $Q\left(\cos\frac{2\pi}{L}x, \sin\frac{2\pi}{L}x, f\right)$, совпадающий с периодической функцией $f(x)$, $f(x + L) = f(x)$, в узлах интерполяции $x_m = \frac{L}{N}m + x_0$ $(m = 0, 1, ..., N - 1)$, можно выбрать так, чтобы он обладал определенными преимуществами перед алгебраическим интерполяционным многочленом, построенным по значениям функции в узлах

$$x_m = \frac{L}{N}m + x_0, \quad m = 0, 1, ..., N - 1.$$

Во-первых, погрешность тригонометрической интерполяции

$$R_N(x, f) = f(x) - Q\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f\right)$$

равномерно стремится к нулю при $N \to 0$, если $f(x)$ имеет хотя бы вторую производную, причем скорость убывания погрешности автоматически учитывает гладкость $f(x)$, т. е. возрастает с ростом числа $r + 1$ производных. Именно, мы докажем что

$$\max_x |R_N(x)| = O\left(\frac{M_{r+1}}{N^{r-1}}\right), \quad M_{r+1} = \max_x \left|\frac{d^{r+1}f(x)}{dx^{r+1}}\right|.$$

Во-вторых, чувствительность тригонометрического интерполяционного многочлена к погрешностям в задании значений f_m в узлах с ростом числа узлов “почти” не возрастает.

Эти два положительных свойства тригонометрической интерполяции (возрастание точности при увеличении гладкости и вычислительную устойчивость) можно придать, как будет показано, и алгебраической интерполяции функций на отрезке за счет специального выбора узлов интерполяции и использования алгебраических многочленов Чебышева, обладающих многими замечательными свойствами.

При первом чтении книги можно ограничиться сказанным о содержании гл. 2 и сразу перейти к гл. 3.

§ 1. Интерполяция периодических функций

Пусть $f(x)$ — периодическая с некоторым периодом L функция:

$$f_N(x+L) \underset{x}{\equiv} f_N(x), \tag{1}$$

заданная на сетке

$$x_m = \frac{L}{N}m + x_0, \quad m = 0, \pm 1, ..., \tag{2}$$

где N — некоторое натуральное число. Для краткости будем обозначать $f(x_m)$ через f_m:

$$f_{m+N} \underset{m}{\equiv} f_m. \tag{3}$$

1. Важный случай выбора узлов интерполяции и соответствующий тригонометрический интерполяционный многочлен.

Теорема 1. *Пусть $x_0 = L/(2N)$, $N = 2(n+1)$, n натуральное. При произвольном задании значений f_m периодической с периодом L функции в узлах сетки (2) существует один и только один интерполяционный тригонометрический многочлен*

$$Q_n\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f\right) = \sum_{k=0}^{n} a_k \cos\frac{2\pi k}{L}x + \sum_{k=1}^{n+1} b_k \sin\frac{2\pi k}{L}x, \tag{4}$$

удовлетворяющий равенствам

$$Q_n|_{x=x_m} = f_m, \quad m = 0, \pm 1, ... \tag{5}$$

Коэффициенты этого многочлена задаются формулами

$$a_0 = \frac{1}{N}\sum_{m=0}^{N-1} f_m, \tag{6}$$

$$a_k = \frac{2}{N}\sum_{m=0}^{N-1} f_m \cos k\left(\frac{2\pi}{N}m + \frac{\pi}{N}\right), \quad k = 1, 2, ..., n, \tag{7}$$

$$b_k = \frac{2}{N}\sum_{m=0}^{N-1} f_m \sin k\Big(\frac{2\pi}{N}m + \frac{\pi}{N}\Big), \quad k = 1, 2, ..., n, \tag{8}$$

$$b_{n+1} = \frac{1}{N}\sum_{m=0}^{N-1} f_m (-1)^m. \tag{9}$$

Доказательство. Рассмотрим множество всех вещественных периодических функций f_m:

$$f_{m+N} = f_m, \quad m = 0, \pm 1, ..., \tag{10}$$

определенных в точках $x_m = -\frac{L}{N}m + \frac{L}{2N}$. Будем рассматривать эти функции только при $m = 0, 1, ..., N-1$, поскольку при остальных m значения этих функций однозначно восстанавливаются по указанным значениям в силу периодичности (10).

Совокупность этих функций с обычными операциями сложения и умножения на вещественные числа образует линейное пространство. Размерность этого пространства, которое обозначим F_N, есть N, поскольку система функций

$$\widetilde{\psi}_m^{(k)} = \begin{cases} 0, & \text{если} \quad m \neq k-1, \\ 1, & \text{если} \quad m = k-1, k = 1, 2, ..., N, \end{cases}$$

образует базис. Действительно, каждую функцию $f \in F_N$, $f = \{f_m\}$ $(m = 0, 1, ..., N-1)$, можно, и притом единственным образом, представить в виде линейной комбинации функций $\widetilde{\psi}_m^{(k)}$. Введем в F_N скалярное умножение

$$(f, g) = \frac{1}{N}\sum_{m=0}^{N-1} f_m g_m.$$

Покажем теперь, что система функций

$$\xi_m^{(0)} = \cos 0 \cdot x_m \equiv 1, \tag{11}$$

$$\xi_m^{(k)} = \sqrt{2}\cos((2\pi k/L)x_m), \quad k = 1, 2, ..., n, \tag{12}$$

$$\eta_m^{(k)} = \sqrt{2}\sin((2\pi k/L)x_m), \quad k = 1, 2, ..., n, \tag{13}$$

$$\eta_m^{(n+1)} = \sin((2\pi(n+1)/L)x_m) = (-1)^m \tag{14}$$

образует ортонормированный базис пространства F_N.

Общее число функций (11)–(14) есть N и равно размерности пространства F_N. Поэтому остается доказать равенства

$$(\xi^{(k)}, \xi^{(k)}) = 1, \quad k = 0, 1, ..., n, \tag{15}$$

$$(\eta^{(k)}, \eta^{(k)}) = 1, \quad k = 1, 2, ..., n+1, \tag{16}$$

$$(\xi^{(r)}, \xi^{(s)}) = 0, \quad r \neq s; \quad r, s = 0, 1, ..., n, \tag{17}$$

$$(\eta^{(r)}, \eta^{(s)}) = 0, \quad r \neq s; \quad r, s = 1, 2, ..., n+1, \tag{18}$$

$$(\xi^{(r)}, \eta^{(s)}) = 0, \quad r = 0, 1, ..., n; \quad s = 1, 2, ..., n+1. \tag{19}$$

Для доказательства равенств (15)–(19) предварительно заметим, что при любых N и γ

$$\frac{1}{N}\sum_{m=0}^{N-1} 1 = 1, \tag{20}$$

$$\sum_{m=0}^{N-1} \cos l\Big(\frac{2\pi}{N}m+\gamma\Big) = 0, \quad l = 1,2,...,N-1, \tag{21}$$

$$\sum_{m=0}^{N-1} \sin l\Big(\frac{2\pi}{N}m+\gamma\Big) = 0, \quad l = 1,2,...,N-1. \tag{22}$$

Действительно, справедливость (20) очевидна. Далее проверим (21):

$$\sum_{m=0}^{N-1} \cos l\Big(\frac{2\pi m}{N}+\gamma\Big) =$$
$$= \frac{1}{2}\sum_{m=0}^{N-1}\Big[\exp\Big\{i\Big(\frac{2l\pi m}{N}+l\gamma\Big)\Big\} + \exp\Big\{-i\Big(\frac{2l\pi m}{N}+l\alpha\Big)\Big\}\Big] =$$
$$= \frac{1}{2}\,e^{il\gamma}\sum_{m=0}^{N-1}\Big(\exp\Big\{i\frac{2l\pi}{N}\Big\}\Big)^m + \frac{1}{2}\,e^{-il\gamma}\sum_{m=0}^{N-1}\Big(\exp\Big\{-i\frac{2l\pi}{N}\Big\}\Big)^m =$$
$$= \frac{1}{2}\,e^{il\gamma}\frac{1-e^{i2l\pi}}{1-\exp\{i2l\pi/N\}} + \frac{1}{2}\,e^{-il\gamma}\frac{1-e^{-i2l\pi}}{1-\exp\{-i2l\pi/N\}} = 0+0 = 0,$$
$$l = 1,2,...,N-1.$$

Равенство (22) доказывается аналогично.

Теперь установим справедливость (15)–(19). Равенство (15) при $k=0$ и равенство (16) при $k=n+1$ совпадают с (20). Если $k = 1,2,...,n$, то (15), (16) справедливы в силу (20), (21):

$$(\xi^{(k)},\xi^{(k)}) =$$
$$= \frac{2}{N}\sum_{m=0}^{N-1}\cos^2\Big(\frac{2k\pi}{N}m+\frac{\pi k}{N}\Big) = \frac{1}{N}\sum_{m=0}^{N-1}\Big[1+\cos\Big(\frac{4k\pi}{N}m+\frac{2\pi k}{N}\Big)\Big] =$$
$$= \frac{1}{N}\sum_{m=0}^{N-1} 1 + \sum_{m=0}^{N-1}\cos 2k\Big(\frac{2\pi}{N}m+\frac{\pi}{N}\Big) = 1+0 = 1,$$

$$(\eta^{(k)},\eta^{(k)}) =$$
$$= \frac{2}{N}\sum_{m=0}^{N-1}\sin^2\Big(\frac{2k\pi}{N}m+\frac{\pi}{N}\Big) = \frac{1}{N}\sum_{m=0}^{N-1}\Big[1-\cos\Big(\frac{2k\pi}{N}m+\frac{2\pi}{N}\Big)\Big] = 1.$$

Докажем (17):

$$(\xi^{(k)}, \xi^{(k)}) = \frac{2}{N}\sum_{m=0}^{N-1} \cos r\Big(\frac{2\pi m}{N} + \frac{\pi}{N}\Big) \cos s\Big(\frac{2\pi m}{N} + \frac{\pi}{N}\Big) =$$

$$= \frac{1}{N}\sum_{m=0}^{N-1} \Big[\cos(r+s)\Big(\frac{2\pi m}{N} + \frac{\pi}{N}\Big) + \cos(r-s)\Big(\frac{2\pi m}{N} + \frac{\pi}{N}\Big)\Big] = 0.$$

Мы воспользовались тем, что $1 \leqslant |r \pm s| \leqslant N-1$ и справедливо (21).

Равенство (18) доказывается аналогично, но вместо тождества $2\cos\alpha\cos\beta = \cos(\alpha+\beta) - \cos(\alpha-\beta)$, использованного при доказательстве (17), нужно использовать тождество $2\sin\alpha\sin\beta = \cos(\alpha - \beta) - \cos(\alpha+\beta)$. Для доказательства (19) используется тождество $2\sin\alpha\cos\beta = \sin(\alpha+\beta) + \sin(\alpha-\beta)$.

Итак, установлено, что (11)–(14) — ортонормированный базис пространства F_N. Поэтому каждая функция $f = \{f_m\} \in F_N$ может быть представлена в виде

$$f_m = \sum_{k=0}^{n} a_k \cos\frac{2\pi k}{L} x_m + \sum_{k=1}^{n+1} b_k \sin\frac{2\pi k}{L} x_m,$$

т. е. в виде линейной комбинации элементов базиса (11)–(14). Умножая сеточные функции, входящие в левую и правую части этого равенства, скалярно на $\xi^{(r)}$ или $\eta^{(s)} (r = 0, 1, ..., n;\ s = 1, 2, ..., n+1)$, получаем равенства

$$\begin{aligned} a_0 &= (f, \xi^{(0)}), \\ a_k &= \sqrt{2}(f, \xi^{(k)}), \quad k = 1, 2, ..., n, \\ b_k &= \sqrt{2}(f, \eta^{(k)}), \quad k = 1, 2, ..., n, \\ b_{n+1} &= (f, \eta^{(n+1)}), \end{aligned}$$

которые совпадают с формулами (6)–(9). □

Сетка $x_m = \frac{L}{N} m + \frac{L}{2N}$ $(m = 0, \pm 1, ...)$, использованная для задания сеточных функций $f \in F_N$ в теореме 1, симметрична относительно точки $x = 0$, так что вместе с точкой $x = x_m$ сетка содержит точку $x = -x_m = x_{-(m+1)}$. Поэтому можно говорить о четных или о нечетных сеточных функциях.

Сеточная функция $f(x)$ $(x = x_m)$ *четная*, если $f(-x) = f(x)$ $(x = x_m)$, или

$$f(x_m) = f(x_{-(m+1)}), \quad m = 0, \pm 1, ... \tag{23}$$

Сеточная функция $f(x)$ $(x = x_m)$ *нечетная*, если $f(-x) = -f(x)$ $(x = x_m)$, или

$$f(x_m) = -f(x_{-(m+1)}), \quad m = 0, \pm 1, ... \tag{24}$$

Теорема 2. *Пусть на сетке* $x_m = \frac{L}{N}m + \frac{L}{2N}$, $N = 2(n+1)$, *задана четная периодическая с периодом* N *сеточная функция* f_m. *Тогда интерполяционный многочлен* (4) *примет вид*

$$Q_n = \sum_{k=0}^{n} a_k \cos\frac{2\pi k}{L}x, \tag{25}$$

причем

$$a_0 = \frac{1}{n+1}\sum_{m=0}^{n} f_m, \tag{26}$$

$$a_k = \frac{2}{n+1}\sum_{m=0}^{n} f_m \cos k\Big(\frac{\pi}{n+1}m + \frac{\pi}{2(n+1)}\Big), \quad k = 1, 2, ..., n. \tag{27}$$

Доказательство. В силу четности (23) из (6), (7) получаем (25), (26), а из (8), (9) следует, что $b_k \equiv 0$. □

Теорема 3. *Пусть на сетке* $x_m = \frac{L}{N}m + \frac{L}{2N}$ *задана нечетная сеточная функция* f_m. *Тогда интерполяционный многочлен* (4) *примет вид*

$$Q_n = \sum_{k=1}^{n+1} b_k \sin\frac{2\pi k}{L}x, \tag{28}$$

$$b_k = \frac{2}{n+1}\sum_{m=0}^{n} f_m \sin\frac{2\pi k}{L}x_m, \quad k = 1, 2, ..., n, \tag{29}$$

$$b_{n+1} = \frac{1}{n+1}\sum_{m=0}^{n} f_m(-1)^m. \tag{30}$$

Доказательство. Благодаря нечетности (24) коэффициенты a_k $(k = 0, 1, ..., n)$ в силу формул (6), (7) окажутся равными нулю, формулы (8), (9) совпадут с (29), (30), а многочлен (4) примет вид (28). □

2. Чувствительность интерполяционного многочлена к погрешностям задания функции в узлах интерполяции. Оценим, какова чувствительность интерполяционного многочлена (4) к погрешности в задании значений f_m. Пусть вместо $f = \{f_m\}$ задана сеточная функция $f + \delta f = \{f_m + \delta f_m\}$. Тогда вместо многочлена (4) получим многочлен

$$Q_n\Big(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f + \delta f\Big) = Q_n\Big(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f\Big) + \delta Q_n.$$

Из формул (6)–(9) видно, что возникающая погрешность есть

$$\delta Q_n = Q_n\Big(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ \delta f\Big).$$

Таким образом, мерой чувствительности интерполяционного многочлена (4) к возмущению δf входных данных могут служить числа

$$L_n = \sup_{f \in F_N} \frac{\max\limits_x \left| Q_n\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f\right)\right|}{\max\limits_m |f_m|}, \quad n = 1, 2, ..., \tag{31}$$

называемые *константами Лебега*. Итак,

$$\max_x |\delta Q_n| \leqslant L_n \max_m |f_m|.$$

Теорема 4. *Константы Лебега L_n тригонометрических интерполяционных многочленов* (4) *удовлетворяют оценкам*

$$L_n \leqslant 2n. \tag{32}$$

Доказательство. Из формул (6), (9) следует, что

$$|a_k| \leqslant 2 \max_m |f_m|, \tag{33}$$

$$|b_k| \leqslant 2 \max_m |f_m|. \tag{34}$$

Из (4), (33) и (34) следует, что

$$\left| Q_n\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f\right)\right| \leqslant \sum_{k=0}^{n} |a_k| + \sum_{k=1}^{n+1} |b_k| \leqslant 2(n+1) \max_m |f_m|. \tag{35}$$

Оценка (35) выполнена для любой $f \in F_N$ и любого N. Отсюда следует (32). □

3. Оценка погрешности интерполяции.

Теорема 5. *Пусть $f(x)$ — периодическая с некоторым периодом L функция, имеющая непрерывную производную некоторого порядка $r+1$:*

$$\max |f^{(r+1)}(x)| = M_{r+1}. \tag{36}$$

Пусть $f \in F_N$, $f = \{f_m\}$ — таблица значений этой функции в точках сетки

$$x_m = \frac{L}{N}m + \frac{L}{2N}, \quad m = 0, \pm 1, ..., \quad N = 2(n+1),$$

а $Q_n\left(\cos\frac{2\pi}{L}x, \sin\frac{2\pi}{L}x, f\right)$ — соответствующий интерполяционный многочлен (4). *Тогда для погрешности интерполяции*

$$R_{n+1}(x) = f(x) - Q_n\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f\right)$$

справедлива оценка

$$|R_{n+1}(x)| \leqslant \frac{CM_{r+1}}{n^{r-1}}, \quad C = C(L) = \text{const}. \tag{37}$$

Доказательство. Проведем предварительные построения. Представим $f(x)$ в виде суммы ряда Фурье

$$f(x) = S_{n+1}(x) + \delta S_{n+1}(x),$$

где

$$S_{n+1}(x) = \sum_{k=0}^{n} \alpha_k \cos\frac{2\pi k}{L}\,x + \sum_{k=1}^{n+1} \beta_k \sin\frac{2\pi k}{L}\,x,$$

$$\delta S_{n+1} = \sum_{k=n+1}^{\infty} \alpha_k \cos\frac{2\pi k}{L}\,x + \sum_{k=n+2}^{\infty} \beta_k \sin\frac{2\pi k}{L}\,x,$$

а α_k, β_k — коэффициенты разложения функции $f(x)$ в ряд Фурье.

Ниже мы используем оценку

$$|\delta S_{n+1}(x)| \leqslant \left(\frac{L}{2\pi}\right)^{r+1} \frac{M_{r+1}}{n^r}.$$

Докажем ее. Из формул

$$\alpha_k = \frac{2}{L}\int_0^L f(x)\cos\frac{2\pi k}{L}\,x\,dx, \qquad \beta_k = \frac{2}{L}\int_0^L f(x)\sin\frac{2\pi k}{L}\,x\,dx,$$

интегрируя их $r+1$ раз по частям, получаем оценки

$$|\alpha_k| \leqslant \left(\frac{L}{2\pi}\right)^{r+1} \frac{2M_{r+1}}{k^{r+1}}, \qquad |\beta_k| \leqslant \left(\frac{L}{2\pi}\right)^{r+1} \frac{2M_{r+1}}{k^{r+1}}.$$

Отсюда

$$\max_x |\delta S_{n+1}(x)| \leqslant \sum_{k=n+1}^{\infty} [|\alpha_k| + |\beta_k|] \leqslant 4\left(\frac{L}{2\pi}\right)^{r+1} \frac{M_{r+1}}{rn^r} = A\frac{M_{r+1}}{n^r}, \quad A = \text{const}. \tag{38}$$

Заметим, что $S_{n+1}(x)$ есть тригонометрический многочлен вида (4). В силу единственности интерполяционного тригонометрического многочлена $Q_n\left(\cos\frac{2\pi}{L}\,x,\ \sin\frac{2\pi}{L}\,x,\ S_{n+1}\right)$ этот многочлен совпадает с $S_{n+1}(x)$:

$$Q_n\left(\cos\frac{2\pi}{L}\,x,\ \sin\frac{2\pi}{L}\,x,\ S_{n+1}\right) = S_{n+1}(x). \tag{39}$$

Далее, в силу оценок (32), (38)

$$\left|Q_n\left(\cos\frac{2\pi}{L}\,x,\ \sin\frac{2\pi}{L}\,x,\ \delta S_{n+1}\right)\right| \leqslant L_n A\,\frac{M_{r+1}}{n^r} \leqslant 2A\,\frac{M_{r+1}}{n^{r-1}}. \tag{40}$$

В силу (38), (40) получим требуемую оценку (37):

$$|R_n(x)| = \left|f(x) - Q_n\left(\cos\frac{2\pi}{L}\,x,\ \sin\frac{2\pi}{L}\,x,\ f\right)\right| =$$

$$= \left| f(x) - Q_n\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ S_{n+1}\right) - \right.$$
$$\left. - Q_n\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ \delta S_{n+1}\right)\right| =$$
$$= \left|(f(x) - S_{n+1}(x)) - Q_n\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ \delta S_{n+1}\right)\right| =$$
$$= \left|\delta S_{n+1}(x) - Q_n\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ \delta S_{n+1}\right)\right| \leqslant$$
$$\leqslant |\delta S_{n+1}(x)| + \left|Q_n\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ \delta S_{n+1}\right)\right| \leqslant$$
$$\leqslant A\frac{M_{r+1}}{n^r} + 2A\frac{M_{r+1}}{n^{r-1}} \leqslant 3A\frac{M_{r+1}}{n^{r-1}}. \quad \square$$

Заметим, что оценки (32) констант Лебега, а также оценки погрешности интерполяции, доказанные нами, можно заменить более сильными.

4. Еще один случай выбора узлов при тригонометрической интерполяции. Отметим еще одну важную для приложений формулу тригонометрической интерполяции, посвятив ей следующую теорему.

Теорема 6. *Пусть $N = 2n$. При произвольном задании значений f_m периодической с периодом L функции в узлах сетки*

$$x_m = \frac{L}{N}m, \quad m = 0, \pm 1, ...,$$

существует один и только один интерполяционный тригонометрический многочлен

$$\widetilde{Q}_n\left(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f\right) = \sum_{k=0}^{n} \widetilde{a}_k \cos\frac{2\pi k}{L}x + \sum_{k=1}^{n-1} \widetilde{b}_k \sin\frac{2\pi k}{L}x, \quad (41)$$

удовлетворяющий равенствам

$$\widetilde{Q}_n\Big|_{x=x_m} = f_m, \quad m = 0, \pm 1, ...$$

Коэффициенты этого многочлена задаются формулами

$$\widetilde{a}_0 = \frac{1}{N}\sum_{m=0}^{N-1} f_m, \quad \widetilde{a}_n = \frac{1}{N}\sum_{m=0}^{N-1} f_m(-1)^m,$$
$$\widetilde{a}_k = \frac{2}{N}\sum_{m=0}^{N-1} f_m \cos\frac{2\pi km}{N}, \quad k = 1, 2, ..., n-1, \quad (42)$$
$$\widetilde{b}_k = \frac{2}{N}\sum_{m=0}^{N-1} f_m \sin\frac{2\pi km}{N}, \quad k = 1, 2, ..., n-1.$$

Доказательство аналогично доказательству теоремы 1, и мы его опускаем.

Отметим, что в случае четной сеточной функции $f(x)$, $f(-x_m) = f(x_m)$, или $f_m = f_{-m}$, формулы (42) принимают вид

$$\widetilde{a}_0 = \frac{1}{N}(f_0 + f_n) + \frac{2}{N}\sum_{m=1}^{n-1} f_m, \quad \widetilde{a}_n = \frac{1}{N}[f_0 + (-1)^n f_n],$$

$$\widetilde{a}_k = \frac{2}{N}[f_0 + (-1)^k f_n] + \frac{4}{N}\sum_{m=1}^{n-1} f_m \cos\frac{2\pi km}{N},$$

$$k = 1, 2, ..., n-1, \quad \widetilde{b}_k \equiv 0, \tag{43}$$

а многочлен (41) принимает вид

$$\widetilde{Q}_n\Big(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f\Big) = \sum_{k=0}^{n} \widetilde{a}_k \cos\frac{2\pi km}{L}.$$

Интерполяционный многочлен $\widetilde{Q}_n\Big(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f\Big)$, задаваемый формулами (41), (42), обладает свойством устойчивости относительно возмущения значений f_m и свойством сходимости при $n \to \infty$ к интерполируемой функции со скоростью, реагирующей на ее гладкость. Эти свойства аналогичны установленным в теоремах 4, 5 для многочлена $Q_n\Big(\cos\frac{2\pi}{L}x,\ \sin\frac{2\pi}{L}x,\ f\Big)$, задаваемого формулой (4).

§ 2. Интерполяция функций на отрезке. Связь между алгебраической и тригонометрической интерполяциями

Пусть $f(x)$ определена на отрезке $-1 \leqslant x \leqslant 1$ и имеет на этом отрезке ограниченную производную некоторого порядка $r+1$.

Мы считаем областью определения функции $f(x)$ отрезок $-1 \leqslant x \leqslant 1$, а не произвольный отрезок $a \leqslant x \leqslant b$, лишь для удобства. Действительно, преобразование $x = \dfrac{t(b-a)+a+b}{2}$ позволяет перейти от функции $f(x)$, определенной на произвольном отрезке $a \leqslant x \leqslant b$, к функции $F(t) \equiv f\Big(\dfrac{t(b-a)+b+a}{2}\Big)$, определенной на отрезке $-1 \leqslant t \leqslant 1$.

1. Периодизация. В силу теоремы 5 из § 1 тригонометрическая интерполяция непосредственно пригодна и эффективна для восстановления лишь гладких периодических функций по их таблицам. Поэтому использование тригонометрической интерполяции для приближенного описания функции $f(x)$ $(-1 \leqslant x \leqslant 1)$ требует перехода от функции $f(x)$ к некоторой гладкой периодической функции. Простое доопределение функции $f(x)$ $(-1 < x \leqslant 1)$ на всей числовой оси до

периодической с периодом $L = 2$ функции приводит, вообще говоря,

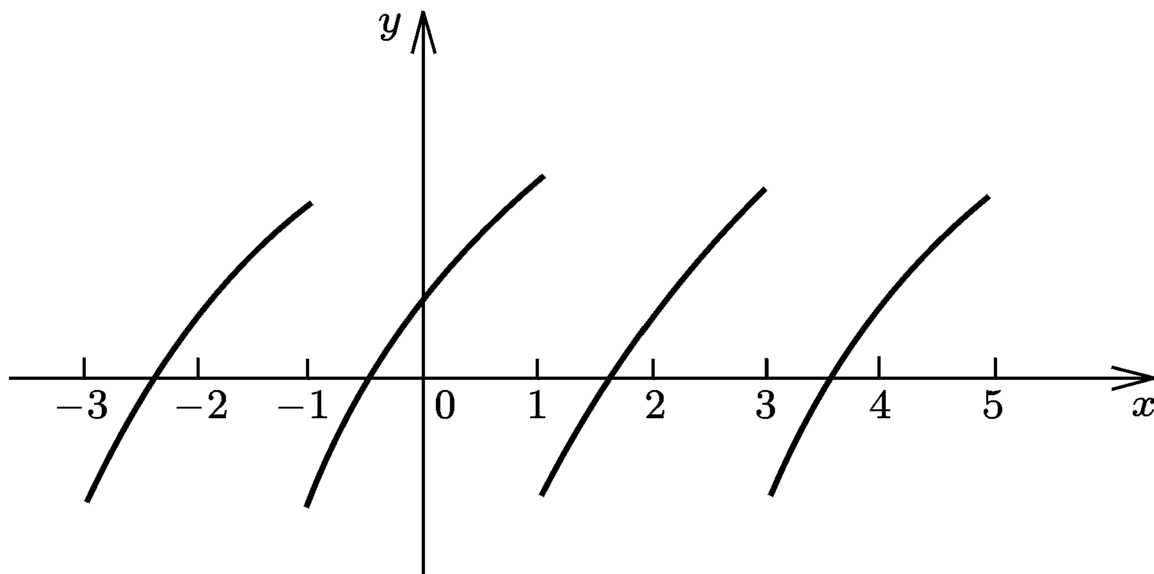


Рис. 3

к разрывной функции (рис. 3). Поэтому перейдем от $f(x)$ к функции

$$F(\varphi) \equiv f(\cos\varphi), \quad x = \cos\varphi. \qquad (1)$$

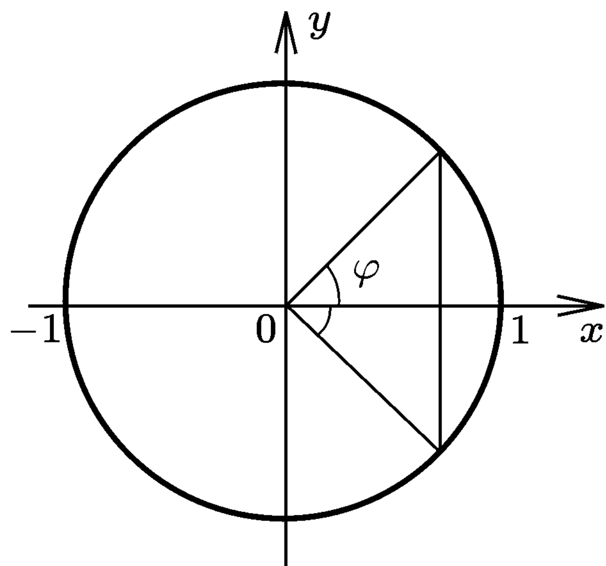


Рис. 4

Для наглядности будем считать, что функция $F(\varphi)$ определена на единичной окружности как функция полярного угла φ. Значение $F(\varphi)$ получается переносом значения $f(x)$ из точки $x \in [-1, 1]$ в точку φ на единичной окружности (рис. 4).

Очевидно, что функция $F(\varphi)$ является четной периодической с периодом 2π функцией. Легко видеть также, что существует и ограничена производная $\dfrac{d^{r+1}}{d\varphi^{r+1}} F(\varphi)$.

2. Тригонометрическая интерполяция. Выберем в качестве узлов интерполяции точки

$$\varphi_m = \frac{2\pi}{N} m + \frac{\pi}{N}, \quad m = 0, \pm 1, ..., \pm m; \quad N = 2(n+1). \qquad (2)$$

Значения $F_m = F(\varphi_m)$ функции $F(\varphi)$ в узлах φ_m в силу определения функции $F(\varphi)$ совпадают со значениями $f_m = f(x_m)$ исходной функции $f(x)$ в точках $x_m = \cos\varphi_m$.

Для интерполяции четной функции $F(\varphi)$ по ее значениям в узлах воспользуемся интерполяционной формулой (25) из § 1

$$Q_n(\cos\varphi, \sin\varphi, F) = \sum_{k=0}^{n} a_k \cos k\varphi, \qquad (3)$$

где коэффициенты a_k в силу $F_m = j_m$ определяются формулами (26),

(27) из § 1:

$$a_0 = \frac{1}{n+1}\sum_{m=0}^{n} f_m,$$
$$a_k = \frac{2}{n+1}\sum_{m=0}^{n} f_m \cos k\varphi_m, \quad k = 1, 2, ..., n. \tag{4}$$

3. Многочлены Чебышева. Связь между тригонометрической и алгебраической интерполяциями. Воспользуемся равенством $\cos\varphi = x$ и введем функции

$$T_k(x) = \cos k\varphi = \cos k \arccos x, \quad k = 0, 1, ... \tag{5}$$

Теорема 1. *Функции $T_k(x)$ суть многочлены степени $k = 0, 1, ...$ При этом $T_0(x) = 1$, $T_1(x) = x$, а $T_2(x), T_3(x), ...$ последовательно вычисляются по рекуррентной формуле*

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x). \tag{6}$$

Доказательство. Очевидно, что $T_0(x) = \cos 0 = 1$, $T_1(x) = \cos\arccos x = x$. Воспользуемся известным тригонометрическим тождеством

$$\cos(k+1)\varphi = 2\cos\varphi\cos k\varphi - \cos(k-1)\varphi, \quad k = 1, 2, ...,$$

которое в случае $\varphi = \arccos x$ переходит в формулу (6).

Докажем, что $T_k(x)$ — многочлен степени k, с помощью индукции по k.

При $k = 0$, $k = 1$ это уже доказано. Фиксируем $k \geqslant 1$. Допустим, что утверждение уже доказано для $T_j(x)$ $(j = 0, 1, ..., k)$, т. е. установлено, что $T_j(x)$ — многочлены степени j $(j = 0, 1, ..., k)$. Тогда выражение в правой части формулы (6), а следовательно, и $T_{k+1}(x)$ есть многочлен степени $k + 1$. □

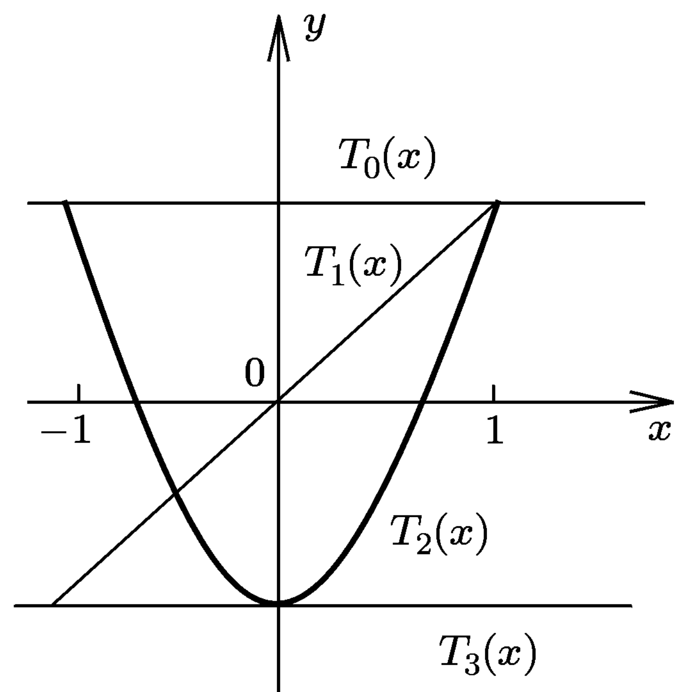


Рис. 5

Многочлены $T_k(x)$ были введены П.Л. Чебышевым. Приведем несколько первых многочленов Чебышева и их графики (рис. 5):

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x.$$

Воспользуемся теперь равенством $\varphi = \arccos x$ и перейдем в (3) от φ к x, положив

$$Q_n(\cos\varphi, \sin\varphi, F) \equiv P_n(x, f),$$

где
$$P_n(x,f)=\sum_{k=0}^{n}a_kT_k(x), \tag{7}$$
$$a_0=\frac{1}{n+1}\sum_{m=0}^{n}f_m=\frac{1}{n+1}\sum_{m=0}^{n}f_mT_0(x_m),$$
$$a_k=\frac{2}{n+1}\sum_{m=0}^{n}f_m\cos k\varphi_m=\frac{2}{n+1}\sum_{m=0}^{n}f_mT_k(x_m). \tag{8}$$

Напомним, что
$$x_m=\cos\varphi_m=\cos\frac{\pi(2m+1)}{n+1},\quad m=0,1,...,n. \tag{9}$$

Таким образом, $P_n(x,f)$ есть алгебраический многочлен степени не выше n, принимающий в узлах интерполяции $x_m=\cos\varphi_m$ заданные значения $f(x_m)=f_m$. Узлы интерполяции x_m изображены на рис. 6.

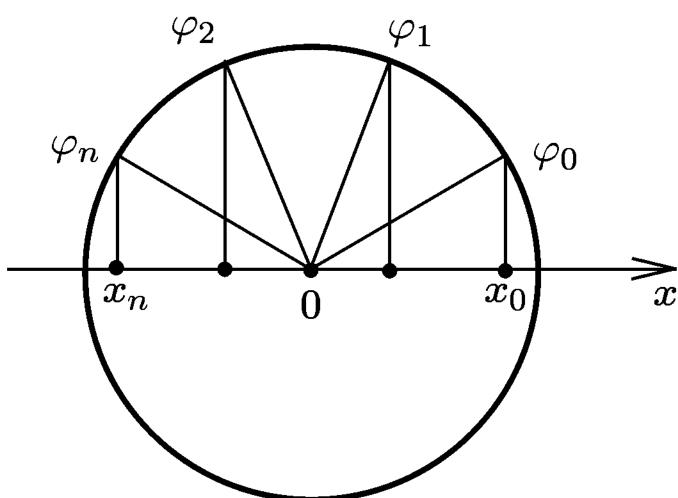


Рис. 6

Отметим, что точки
$$\varphi_m=\frac{\pi}{n+1}m+\frac{\pi}{2(n+1)}$$
$(m=0,1,...,n)$ являются нулями функции $\cos(n+1)\varphi$, а $x_m=\cos\varphi_m$ — нулями многочлена Чебышева $T_{n+1}(x)=\cos(n+1)\varphi$.

Таким образом, интерполяционный многочлен $P_n(x,f)$, задаваемый формулой (7), реализует алгебраическую интерполяцию функции $f(x)$ по ее значениям в точках x_m, являющихся нулями многочлена Чебышева $T_{n+1}(x)$.

4. Свойства алгебраической интерполяции с нулями многочлена Чебышева $T_{n+1}(x)$ в качестве узлов. В силу равенства $Q_n(\cos\varphi,\ \sin\varphi,\ F)=P_n(x,f)$ на алгебраический интерполяционный многочлен $P_n(x,f)$ переносятся свойства многочлена $Q_n(\cos\varphi,\ \sin\varphi,\ F)$, установленные в теоремах 4, 5 из § 1. Именно константы Лебега L_n, характеризующие чувствительность многочлена $P_n(x,f)$ к погрешностям при задании значений f_m, удовлетворяют доказанной оценке (32) из § 1
$$L_n\leqslant 2n, \tag{10}$$
а погрешность интерполяции
$$R_n(x)=f(x)-P_n(x,f)$$
равномерно стремится к нулю при $n\to\infty$ со скоростью, зависящей

от числа производных $r+1$:

$$\max_{-1\leqslant x\leqslant 1}|R_n(x)| = O\left(\frac{M_{r+1}}{n^{r-1}}\right), \quad M_{r+1} = \max_x |f^{(r+1)}(x)|. \qquad (11)$$

Напомним, что при использовании вместо сетки $x_m = \cos\varphi_m$ сетки с постоянным шагом константы Лебега быстро растут, а сходимость интерполяционного многочлена к функции $f(x)$ может не иметь места даже для бесконечно дифференцируемых функций. Это и заставляет в случае равномерной или произвольной нерегулярной сетки использовать кусочно многочленную алгебраическую интерполяцию или сплайны (см. гл. 1).

Замечание. Доказанная нами в теореме 4 из § 1 оценка (32) может быть существенно усилена: справедлива оценка

$$L_n = \frac{2}{\pi}\ln n + 1 - \theta_n, \quad 0 \leqslant \theta_n \leqslant \frac{1}{4}$$

(см., например, [1]). Вместе с тем может быть усилено и (11).

5. Алгоритм вычисления значений интерполяционного многочлена. Для вычисления коэффициентов многочлена (7) по формулам (8), а также для вычисления значения этого многочлена в заданной точке x $(-1 \leqslant x \leqslant 1)$ нужно уметь вычислять значения многочленов $T_k(x)$ $(k = 0, 1, ...)$. Покажем, что эти значения целесообразно вычислять по формуле (6). Простота и экономичность формулы (6) очевидны. Покажем, что вычисления по ней устойчивы относительно погрешностей округления.

Рассмотрим так называемое разностное уравнение вида

$$y_{k+1} = 2xy_k - y_{k-1}, \qquad (12)$$

где k — параметр. Будем искать решение этого уравнения, имеющее вид $y_k = q^k$, где q — некоторое число. Подставляя это выражение в разностное уравнение, получаем следующее уравнение ("характеристическое уравнение") для q:

$$q^2 - 2xq + 1 = 0, \quad q_{1,2} = x \pm \sqrt{x^2-1}. \qquad (13)$$

В силу линейности уравнения (12) выражение

$$y_k = c_1 q_1^k + c_2 q_2^k \qquad (14)$$

является его решением при любых значениях постоянных c_1, c_2. Подберем c_1, c_2 из условий

$$y_0 = T_0(x) = 1, \quad y_1 = T_1(x) = x,$$

т. е.

$$c_1 + c_2 = 1, \quad c_1 q_1 + c_2 q_2 = x. \qquad (15)$$

Отсюда при $c_1 = c_2 = 1/2$ решение (14) совпадает с $T_k(x)$, которое в силу соотношения (6) является решением уравнения (12) как функция k:

$$T_k(x) = \frac{1}{2}(x + \sqrt{x^2-1})^k + \frac{1}{2}(x - \sqrt{x^2-1})^k. \qquad (16)$$

Заметим, что корни $q_{1,2} = x \pm \sqrt{x^2-1}$ при $|x| < 1$ являются комплексно-сопряженными и по модулю равны единице. Следовательно, q_1^k, q_2^k с ростом k по модулю не меняются и остаются равными единице.

Какая-либо погрешность, допущенная при некотором $k = k_0$, вызовет возмущение в значениях c_1, c_2, которые входят в формулу (14) при $k > k_0$. В силу равенств $|q_1^k| = |q_2^k| = 1$ $(k = 1, 2, ...)$ эта погрешность с ростом k не будет возрастать, что и означает вычислительную устойчивость расчетов по формуле (6) при $|x| < 1$.

6. Алгебраическая интерполяция с узлами в точках экстремума многочлена Чебышева $T_n(x)$. При интерполяции функции $F(\varphi) = f(\cos\varphi)$ воспользуемся сеткой $\widetilde{\varphi}_m = \frac{\pi}{n} m$ $(m = 0, 1, ...$ $..., 2n-1)$. В соответствии с теоремой 6 из § 1 получим тригонометрический интерполяционный многочлен

$$\widetilde{Q}_n(\cos\varphi, \sin\varphi, F) = \sum_{k=0}^{n} \widetilde{a}_k \cos k\varphi,$$

$$\widetilde{a}_0 = \frac{1}{2n}(f_0 + f_n) + \frac{1}{n}\sum_{m=1}^{n-1} f_m, \quad \widetilde{a}_n = \frac{1}{2n}[f_0 + (-1)^n f_n],$$

$$\widetilde{a}_k = \frac{1}{n}[f_0 + (-1)^k f_n] + \frac{2}{n}\sum_{m=1}^{n-1} f_m \cos k \cdot \varphi_m,$$

$$k = 1, 2, ..., n-1.$$

Переходя к переменной $x = \cos\varphi$ и обозначая $\widetilde{Q}_n(\cos\varphi, \sin\varphi, F) =$ $= \widetilde{P}_n(x, f)$, получаем

$$\widetilde{P}_n(x, f) = \sum_{k=0}^{n} \widetilde{a}_k T_k(x), \tag{17}$$

$$\begin{gathered} \widetilde{a}_0 = \frac{1}{2n}(f_0 + f_n) + \frac{1}{n}\sum_{m=1}^{n-1} f_m, \quad \widetilde{a}_n = \frac{1}{2n}[f_0 + (-1)^n f_n], \\ \widetilde{a}_k = \frac{1}{n}[f_0 + (-1)^k f_n] + \frac{2}{n}\sum_{m=1}^{n-1} f_m T_k(x_m). \end{gathered} \tag{18}$$

Алгебраический интерполяционный многочлен $\widetilde{P}_n(x, f)$ с узлами интерполяции

$$\widetilde{x}_m = \cos\varphi_m = \cos\frac{\pi}{n}m, \quad m = 0, 1, ..., n,$$

наследует от тригонометрического интерполяционного многочлена $\widetilde{Q}_n(\cos\varphi, \sin\varphi, F)$ слабый рост констант Лебега с увеличением n (устойчивость относительно погрешности значений f_m), а также увеличение порядка малости погрешности интерполяции относительно

числа $1/n$ при возрастании гладкости функции $f(x)$.

Обратим внимание читателя на тот очевидный факт, что в узлах интерполяции $\widetilde{x}_m$ многочлен Чебышева $T_n(x)$ достигает своих экстремальных на отрезке $-1 \leqslant x \leqslant 1$ значений: $T_n(\widetilde{x}_m) = \cos \pi m = (-1)^m (m = 0, 1, ..., n)$, что и оправдывает заголовок этого пункта.

Задачи

1. Пусть функция $f(x)$ задана не на отрезке $[-1, 1]$, а на произвольном отрезке $[a, b]$.

Указать узлы интерполяции и выписать интерполяционные многочлены, аналогичные многочленам $P_n(x, f)$, $\widetilde{P}_n(x, f)$, построенным в этом параграфе.

2. Для функции $f(x)$, $-1 \leqslant x \leqslant 1$, построить интерполяционный многочлен $P_n(x)$, используя в качестве узлов интерполяции нули многочлена Чебышева $T_{n+1}(x)$. Вывести на экран компьютера графики $f(x)$ и $P_n(x)$ для $n = 5, 10, 20, 30$. Сделать то же самое в случае интерполяционного многочлена $P_n(x)$, построенного по равноотстоящим узлам $x_k = -1 + 2k/n$, $k = 0, 1, ..., n$. Рассмотреть случай $f(x) = \dfrac{1}{x^2 + 0{,}25}$. Объяснить качественное различие между обоими способами интерполяции, наблюдаемое на экране.

3. Введем нормированный многочлен Чебышева $T_n(x)$ степени n, положив $T_n(x) = 2^{1-n} T_n(x)$.

а) Показать, что коэффициент при x^n для $T_n(x)$ равен единице.

б) Показать, что уклонение $\max\limits_{-1 \leqslant x \leqslant 1} |T_n(x)|$ многочлена $T_n(x)$ от нуля на отрезке $[-1, 1]$ есть 2^{1-n}.

в*) Показать, что среди всех многочленов степени n с коэффициентом 1 при x^n многочлен $T_n(x)$ наименее уклоняется от нуля на отрезке $[-1, 1]$.

г) Как выбрать узлы интерполяции $t_0, t_1, ..., t_n$ на отрезке $[-1, 1]$, чтобы многочлен $(t - t_0)(t - t_1)...(t - t_n)$, входящий в формулу погрешности (2) из § 2 гл. 1, наименее уклонялся от нуля на отрезке $[-1, 1]$?

4. Указать узлы интерполяции четной периодической функции $f(\varphi)$, $f(\varphi + 2\pi) \equiv f(\varphi)$, при которых константы Лебега совпадают с константами Лебега для алгебраической интерполяции по равноотстоящим узлам.

ГЛАВА 3

ВЫЧИСЛЕНИЕ ОПРЕДЕЛЕННЫХ ИНТЕГРАЛОВ. КВАДРАТУРЫ

Определенный интеграл $\int\limits_a^b f(x)\,dx$ удается найти точно по формуле Ньютона–Лейбница лишь в тех редких случаях, когда соответствующий неопределенный интеграл от заданной функции $f(x)$ является

табличным. Еще реже удается находить точные значения кратных интегралов $\iint\limits_D f(x,y)\,dxdy$, где D — заданная область.

Мы укажем некоторые способы приближенного вычисления определенных и кратных интегралов, обсудим трудности, связанные с ростом размерности кратного интеграла.

Отдельный параграф посвящен комбинированному использованию аналитических и вычислительных средств при решении задач, которое широко применяется не только при вычислении интегралов.

Формулы для приближенного вычисления интеграла по таблице значений подынтегральной функции называют *квадратурными* в случае интегралов по отрезку и *кубатурными* в случае кратных интегралов.

§ 1. Квадратурные формулы трапеций и Симпсона

Определенный интеграл $\int\limits_a^b f(x)\,dx$, как известно, имеет геометрический смысл площади.

1. Общая схема простейших квадратурных формул. Для приближенного вычисления этой площади разобьем отрезок $[a,b]$ на некоторое число n отрезков, обозначив абсциссы их концов $x_0, x_1, \ldots$ $\ldots, x_n$, причем

$$a = x_0 < x_1 < \ldots < x_n = b.$$

Затем воспользуемся кусочно многочленной интерполяцией какой-либо степени $k \geqslant 1$, заменив функцию $y = f(x)$ возникающей при этом кусочно многочленной функцией

$$P_k(x) = P_k(x, f, x_0, x_1, \ldots, x_n).$$

Функция $P_k(x)$ на каждом отрезке $[x_i, x_{i+1}]$ есть многочлен степени не выше k.

Для вычисления интеграла можно воспользоваться приближенным равенством

$$\int\limits_a^b f(x)dx \approx \int\limits_a^b P_k(x, f, x_0, x_1, \ldots, x_n)\,dx. \tag{1}$$

Правую часть этого равенства можно записать в виде

$$\int\limits_a^b P_k\,dx = \int\limits_a^{x_1} P_k\,dx + \int\limits_{x_1}^{x_2} P_k\,dx + \ldots + \int\limits_{x_{n-1}}^{b} P_k\,dx. \tag{2}$$

Каждое слагаемое в правой части — интеграл от многочлена степени не выше k, с которым совпадает функция $P_k = P_k(x, f, x_0, x_1, ..., x_n)$ на отрезке $[x_i, x_{i+1}]$. Его можно вычислить точно по формуле Ньютона–Лейбница, выразив ответ через числа $x_0, x_1, ..., x_n$ и $f(x_0), f(x_1), ...$ $..., f(x_n)$. Формула (1) является, таким образом, квадратурной формулой.

Очевидно, справедлива следующая оценка погрешности квадратурной формулы:

$$\left|\int_a^b f(x)\,dx - \int_a^b P_k\,dx\right| \leqslant \int_a^b |f(x) - P_k|\,dx \leqslant$$
$$\leqslant (b-a) \max_{a\leqslant x\leqslant b} |f(x) - P_k(x, f, x_0, x_1, ..., x_n)|. \quad (3)$$

Теорема 1. *Пусть $f(x)$ имеет ограниченную производную порядка $k+1$, так что $\max\limits_{a\leqslant x\leqslant b} |f^{(k+1)}(x)| = M_{k+1}$, и пусть $h = \max\limits_{0\leqslant i\leqslant n-1} |x_{i+1} - x_i|$. Тогда*

$$\left|\int_a^b f(x)\,dx - \int_a^b P_k\,dx\right| \leqslant \mathrm{const}\cdot|b-a|M_{k+1}h^{k+1}. \quad (4)$$

Доказательство. Достаточно воспользоваться следующей оценкой погрешности кусочно многочленной интерполяции:

$$\max_{a\leqslant x\leqslant b} |f(x) - P_k| \leqslant \mathrm{const}\cdot M_{k+1}h^{k+1},$$

а также оценкой (3). □

Квадратурные формулы вида (1) наиболее часто употребляются при использовании кусочно линейной интерполяции $k = 1$ (формула трапеции) и при кусочно квадратичной интерполяции (формула Симпсона).

2. Формула трапеций. Конкретизируем формулу (1) в случае $k = 1$ и при равноотстоящих узлах $x_i, x_{i-1} - x_i = (b-a)/n = h$. В этом случае интеграл $\int\limits_{x_i}^{x_{i+1}} P_1\,dx$ есть площадь трапеции и равен произведению полусуммы оснований этой трапеции $(f_i + f_{i+1})/2$ на ее высоту $h = (b-a)/n$, т. е.

$$\int_{x_i}^{x_{i+1}} P_1(x)\,dx = \frac{b-a}{n}\,\frac{f_i + f_{i+1}}{2}. \quad (5)$$

Суммировав эти равенства почленно для $i = 0, 1, ..., n-1$, придадим

формуле (1) следующий вид:

$$\int_a^b f(x)\,dx \approx \int_a^b P_1 dx = \frac{b-a}{n}\left(\frac{f_0}{2} + f_1 + \ldots + f_{n-1} + \frac{f_n}{2}\right). \qquad (6)$$

Уточним общую оценку (4) погрешности квадратурной формулы (1) для случая формулы трапеций (6). На каждом отрезке $[x_i, x_{i+1}]$ функция $P_1(x)$ совпадает с интерполяционным многочленом первой степени. Поэтому на каждом отрезке $[x_i, x_{i+1}]$ справедлива оценка

$$|f(x) - P_1(x)| \leqslant \frac{1}{8} \max_{x_i \leqslant x \leqslant x_{i+1}} |f''(x)| h^2 \leqslant \frac{1}{8} \max_{a \leqslant x \leqslant b} |f''(x)| h^2.$$

Правая часть этой оценки не зависит от того, какому из отрезков $[x_i, x_{i+1}]$ принадлежит точка x. Поэтому выписанная оценка справедлива при всех $x \in [a, b]$:

$$\max_{a \leqslant x \leqslant b} |f(x) - P_1(x)| \leqslant \frac{1}{8} \max_{a \leqslant x \leqslant b} |f''(x)| h^2. \qquad (7)$$

Оценка погрешности (3) при $k = 1$ и при равноотстоящих узлах x_i, т. е. для формулы трапеций, принимает вид

$$\left|\int_a^b f(x)\,dx - \int_a^b P_1(x, f, x_0, x_1, \ldots, x_n)\,dx\right| \leqslant \frac{b-a}{8} \max_{a \leqslant x \leqslant b} |f''(x)| h^2. \qquad (8)$$

Таким образом, эта оценка гарантирует убывание погрешности при $h \to 0$ не медленнее, чем h^2. Она не является грубой: даже для функций, имеющих производные всех порядков, погрешность квадратурной формулы трапеций, вообще говоря, действительно убывает именно как h^2.

Это показывает следующий пример. Очевидно, что $\int_0^1 x^2 dx = \frac{1}{3}$. По формуле трапеций имеем

$$\int_0^1 x^2 dx \approx \frac{1}{n}\left(\frac{x_0^2}{2} + x_1^2 + \ldots + x_{n-1}^2 + \frac{x_n^2}{2}\right) =$$

$$= \frac{h^2}{n}\left[(1^2 + 2^2 + \ldots + n^2) - \frac{n^2}{2}\right] = \frac{h^2}{n}\left[\frac{n(n+1)(2n+1)}{6} - \frac{n^2}{2}\right] = \frac{1}{3} + \frac{h^2}{6}.$$

Погрешность здесь составляет $h^2/6$, хотя подынтегральная функция $f(x) = x^2$ имеет производные не только второго, но и всех порядков.

Интересно напомнить тот известный факт, что для периодических функций $f(x)$ с периодом $b - a$ точность формулы трапеций для $\int_a^b f(x)\,dx$ самонастраивается на гладкость; порядок точности по h возрастает с ростом числа производных периодической функции:

Теорема 2. *Пусть $f(x)$ — периодическая функция с периодом $X = b - a$, $f(x + X) \equiv f(x)$, имеющая ограниченные производные до некоторого порядка k, $\max\limits_x |f^{(k)}(x)| = M_k$. Тогда погрешность формулы трапеций будет $O(M_k h^k)$.*

Мы не приводим доказательства этой теоремы.

3. Формула Симпсона. Эта квадратурная формула получается на основе замены подынтегральной функции кусочно квадратичной. Для достаточно гладких функций ее погрешность убывает при $h \to 0$ как h^4, т. е. быстрее, чем для формулы трапеций.

Будем считать n четным, $n = 2k$, а узлы равноотстоящими, $x_{j+1} - x_j = h = (b-a)/n$. На каждом отрезке $[x_{2j}, x_{2j+2}]$ $(j = 0, 1, ..., k-1)$ заменим $f(x)$ интерполяционным многочленом второй степени $P_2(x, f_{2j}, f_{2j+1}, f_{2j+2})$, который запишем в форме Лагранжа:

$$P_2(x, f_{2j}, f_{2j+1}, f_{2j+2}) = f_{2j}\frac{(x - x_{2j+1})(x - x_{2j+2})}{(x_{2j} - x_{2j+1})(x_{2j} - x_{2j+2})} + \\ + f_{2j+1}\frac{(x - x_{2j})(x - x_{2j+2})}{(x_{2j+1} - x_{2j})(x_{2j+1} - x_{2j+2})} + f_{2j+2}\frac{(x - x_{2j})(x - x_{2j+1})}{(x_{2j+2} - x_{2j})(x_{2j+2} - x_{2j+1})}.$$

По формуле Ньютона–Лейбница проверим, что

$$\int\limits_{x_{2j}}^{x_{2j+2}} P_2\,dx = \frac{b-a}{3n}(f_{2j} + 4f_{2j+1} + f_{2j+2}). \tag{9}$$

В результате приближенной замены $f(x)$ на $P_2(x, f_{2j}, f_{2j+1}, f_{2j+2})$ получим приближенные равенства

$$\int\limits_{x_{2j}}^{x_{2j+2}} f(x)\,dx \approx \int\limits_{x_{2j}}^{x_{2j+2}} P_2\,dx = \frac{b-a}{3n}(f_{2j} + 4f_{2j+1} + f_{2j+2}), \\ j = 0, 1, ..., k-1.$$

Суммируя приближенные равенства

$$\int\limits_{x_{2j}}^{x_{2j+2}} f(x)\,dx \approx \frac{b-a}{3n}(f_{2j} + 4f_{2j+1} + f_{2j+2}) \tag{10}$$

для $j = 0, 1, ..., k-1$ почленно, в левой части получаем $\int\limits_a^b f(x)\,dx$:

$$\int\limits_{x_0}^{x_2} f(x)\,dx + \int\limits_{x_2}^{x_4} f(x)\,dx + ... + \int\limits_{x_{2n-2}}^{x_{2n}} f(x)\,dx = \int\limits_a^b f(x)\,dx.$$

В правой части получаем

$$S_n(f) = \frac{b-a}{3n}(f_0 + 4f_1 + 2f_2 + 4f_3 + \ldots + 4f_{n-1} + f_n). \qquad (11)$$

Формула (11) для приближенного вычисления определенного интеграла $\int\limits_a^b f(x)\,dx$ называется *формулой Симпсона.*

Теорема 3. *Пусть $f(x)$ имеет на отрезке $[a,b]$ ограниченную третью производную: $\max\limits_{a\leqslant x\leqslant b} |f^{(3)}(x)| = M_3$. Тогда погрешность формулы Симпсона, т. е. разность $\int\limits_a^b f(x)\,dx - S_n(f)$, удовлетворяет оценке*

$$\left|\int\limits_a^b f(x)\,dx - S_n(f)\right| \leqslant \frac{(b-a)M_3}{12}\,h^3.$$

Доказательство. Формула Симпсона возникла в результате замены подынтегральной функции $f(x)$ на отрезке $[a,b]$ кусочно квадратичной функцией $P_2(x)$. Она на каждом отрезке $[x_{2j}, x_{2j+2}]$ совпадает с интерполяционным многочленом $P_2(x, x_{2j}, x_{2j+1}, x_{2j+2})$, построенным по значениям $f_{2j}, f_{2j+1}, f_{2j+2}$ в узлах $x_{2j}, x_{2j+1}, x_{2j+2}$.

Оценим разность $R_2(x) \equiv f(x) - P_2(x)$ для $x \in [a,b]$. Каждый из этих x принадлежит какому-то отрезку $[x_{2j}, x_{2j+2}]$. Но на отрезке $[x_{2j}, x_{2j+2}]$ величина $R_2(x)$ есть погрешность квадратичной интерполяции и выражается формулой

$$R_2(x) = \frac{f^{(3)}(\xi)}{3!}(x - x_{2j})(x - x_{2j+1})(x - x_{2j+2}), \quad \xi \in (x_{2j}, x_{2j+2}).$$

Отсюда

$$|R_2(x)| \leqslant \frac{M_3}{3!} \max_{x_{2j}\leqslant x\leqslant x_{2j+2}} |(x - x_{2j})(x - x_{2j+1})(x - x_{2j+2})| \leqslant \frac{M_3}{12}h^3.$$

В силу произвольности $x \in [a,b]$

$$\max_{a\leqslant x\leqslant b} |R_2(x)| \leqslant \frac{M_3}{12}h^3.$$

Следовательно,

$$\left|\int\limits_a^b f(x)\,dx - S_n(f)\right| = \left|\int\limits_a^b f(x)\,dx - \int\limits_a^b P_2(x)\,dx\right| =$$

$$= \left|\int\limits_a^b (f(x) - P_2(x))\,dx\right| = \left|\int\limits_a^b R_2(x)\,dx\right| \leqslant \int\limits_a^b |R_2(x)|dx \leqslant$$

$$\leqslant \int\limits_a^b \max_x |R_2(x)|dx \leqslant \frac{M_3h^3}{12}\int\limits_a^b dx = \frac{b-a}{12}\,h^3M^3. \quad \square$$

Теорема 4. *Пусть $f(x)$ имеет на отрезке $[a,b]$ ограниченную четвертую производную: $\max\limits_{a \leqslant x \leqslant b} |f^{(4)}(x)| = M_4$. Тогда погрешность формулы Симпсона, т. е. разность $\int\limits_a^b f(x)\,dx - S_n(f)$, удовлетворяет оценке*

$$\Bigl|\int\limits_a^b f(x)\,dx - S_n(f)\Bigr| \leqslant \frac{(b-a)M_4}{18}\,h^4. \tag{12}$$

Доказательство. Формула Симпсона возникла из записи интеграла $\int\limits_a^b f(x)\,dx$ в виде

$$\int\limits_a^b f(x)\,dx = \int\limits_{x_0}^{x_2} f(x)\,dx + \int\limits_{x_2}^{x_4} f(x)\,dx + \ldots + \int\limits_{x_{n-2}}^{x_n} f(x)\,dx \tag{13}$$

с дальнейшей заменой каждого слагаемого $\int\limits_{x_{2j}}^{x_{2j+2}} f(x)\,dx$ выражением $\dfrac{b-a}{3n}(f_{2j} + 4f_{2j+1} + f_{2j+2})$.

Покажем, что погрешность, вносимая в каждое из $k = n/2$ слагаемых суммы (13), удовлетворяет при этом одной и той же оценке

$$\Bigl|\int\limits_{x_{2j}}^{x_{2j+2}} f(x)\,dx - \frac{b-a}{3n}(f_{2j} + 4f_{2j+1} + f_{2j+2})\Bigr| \leqslant \frac{M_4}{9}\,h^5. \tag{14}$$

Из этих оценок вытекает и оценка (12):

$$\Bigl|\int\limits_a^b f(x)\,dx - S_n(f)\Bigr| = \Bigl|\sum_{j=0}^{k-1} \int\limits_{x_{2j}}^{x_{2j+2}} f(x)\,dx - \frac{b-a}{3n}(f_{2j} + 4f_{2j+1} + f_{2j+2})\Bigr| \leqslant$$

$$\leqslant \sum_{j=0}^{k-1} \Bigl|\int\limits_{x_{2j}}^{x_{2j+2}} f(x)\,dx - \frac{b-a}{3n}(f_{2j} + 4f_{2j+1} + f_{2j+2})\Bigr| \leqslant$$

$$\leqslant \frac{n}{2}\frac{M_4}{9}\,h^5 = \frac{(b-a)M_4}{18}\,h^4.$$

Для доказательства оценки (14) представим $f(x)$ с помощью формулы Тейлора в окрестности точки x_{2j+1} в виде

$$f(x) = Q(x) + R(x),$$

где

$$Q(x) = f(x_{2j+1}) + \frac{f'(x_{2j+1})}{1!}(x - x_{2j+1}) +$$

$$+\frac{f''(x_{2j+1})}{2!}(x-x_{2j+1})^2+\frac{f'''(x_{2j+1})}{3!}(x-x_{2j+1})^3,$$

$$R(x)=\frac{f^{(4)}(\xi)}{4!}(x-x_{2j+1})^4, \quad x_{2j}\leqslant\xi\leqslant x_{2j+2}.$$

Далее,

$$\int\limits_{x_{2j}}^{x_{2j+2}} f(x)\,dx-\frac{b-a}{3n}(f_{2j}+4f_{2j+1}+f_{2j+2})=\int\limits_{x_{2j}}^{x_{2j+2}} Q(x)\,dx+\int\limits_{x_{2j}}^{x_{2j+2}} R(x)\,dx-$$

$$-\frac{b-a}{3n}[(Q_{2j}+R_{2j})+4(Q_{2j+1}+R_{2j+1})+(Q_{2j+2}+R_{2j+2})]=$$

$$=\Big\{\int\limits_{x_{2j}}^{x_{2j+2}} Q(x)\,dx-\frac{b-a}{3n}(Q_{2j}+4Q_{2j+1}+Q_{2j+2})\Big\}+$$

$$+\Big[\int\limits_a^b R(x)\,dx-\frac{b-a}{3n}(R_{2j}+4R_{2j+1}+R_{2j+2})\Big].$$

Можно проверить непосредственно, что для $Q(x)\equiv 1$, $Q(x)\equiv x$, $Q(x)\equiv x^2$, $Q(x)\equiv x^3$ величина, стоящая в фигурных скобках, обращается в нуль. Следовательно, эти фигурные скобки обращаются в нуль и для произвольного многочлена $Q(x)$ степени не выше третьей. В силу этого

$$\Big|\int\limits_{x_{2j}}^{x_{2j+2}} f(x)\,dx-\frac{b-a}{3n}(f_{2j}+4f_{2j+1}+f_{2j+2})\Big|=$$

$$=\Big|\int\limits_{x_{2j}}^{x_{2j+2}} R(x)\,dx-\frac{b-a}{3n}(R_{2j}+4R_{2j+1}+R_{2j+2})\Big|. \quad (15)$$

Учтем, что $R_{2j+1}=0$, а $|R(x)|\leqslant\dfrac{M_4}{4!}\,h^4$. Поэтому правая часть равенства (15) оценивается величиной

$$\int\limits_{x_{2j}}^{x_{2j+2}} |R(x)|dx+\frac{b-a}{3n}\cdot 2\max|R(x)|\leqslant\max_x|R(x)|\Big(\int\limits_{x_{2j}}^{x_{2j+2}} dx+\frac{2}{3}\,\frac{b-a}{n}\Big)=$$

$$=\max_x|R(x)|\Big(2\,\frac{b-a}{n}+\frac{2}{3}\,\frac{b-a}{n}\Big)\leqslant\frac{M_4}{4!}\,\frac{8}{3}(b-a)\,\frac{1}{n}\,h^4=\frac{M_4}{9}\,h^5. \quad \square$$

С л е д с т в и е. *Формула Симпсона точна для функций $f(x)$, являющихся многочленами степени не выше третьей.*

Д о к а з а т е л ь с т в о. Для таких функций $M_4=\max|f^{(4)}(x)|=0$ и в правой части оценки (14) стоит нуль. □

Заметим, что на практике выбор числа n и соответственно шага сетки перепоручают машине. Ведут расчет с некоторым n, затем с

удвоенным n, пока результат не перестанет меняться в требуемом десятичном знаке.

Задачи

1. Показать, что формула трапеций точна, если $f(x)$ — многочлен степени не выше первой.

2*. Доказать теорему 2 в случае $b - a = 1$.

Указание. Воспользоваться представлением $f(x)$ в виде ряда Фурье

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{2k\pi x i}.$$

3. Показать, что в случае переменного шага сетки, удовлетворяющего условию $x_{j+1} - x_j < h = \text{const}$, использование кусочно квадратической интерполяции приведет к обобщению формулы Симпсона, порядок погрешности которой возрастает и будет $O(M_3 h^3)$, где $M_3 = \max|f'''(x)|$.

4. Показать, что для сколь угодно гладких подынтегральных функций погрешность формулы Симпсона может фактически достигать величины $\text{const} \cdot h^4$.

§ 2. Сочетание численных и аналитических методов при вычислении интегралов с особенностями

Даже для простейших несобственных интегралов непосредственное применение квадратурных формул наталкивается на трудности. Например, формула трапеций

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}\left(\frac{f_0}{2} + f_1 + \ldots + f_{n-1} + \frac{f_n}{2}\right)$$

в случае $a = 0$, $b = 10$, $f(x) = \cos x/\sqrt{x}$ неприменима, так как f_0 не определено.

Для преодоления этих трудностей естественно воспользоваться аналитическими методами, чтобы свести задачу к вычислению определенного интеграла от гладкой ограниченной функции, а затем воспользоваться какой-либо квадратурной формулой. В рассматриваемом примере с этой целью можно сначала воспользоваться формулой интегрирования по частям, приняв $u = \cos x$, $\dfrac{1}{\sqrt{x}}\,dx = dv$. Тогда

$$\int_0^{10} \frac{\cos x}{\sqrt{x}}\,dx = \cos x\,(2\sqrt{x})\Big|_0^{10} + \int_0^{10} 2\sqrt{x}\sin x\,dx,$$

и осталось вычислить следующий определенный интеграл:

$$\int_0^{10} (\sin x)\sqrt{x}\,dx.$$

Возможен другой прием. А именно, разобьем интеграл $\int\limits_0^{10} \frac{\cos x}{\sqrt{x}}\,dx$ на два:

$$\int\limits_0^{10} \frac{\cos x}{\sqrt{x}}\,dx = \int\limits_0^{c} \frac{\cos x}{\sqrt{x}}\,dx + \int\limits_c^{10} \frac{\cos x}{\sqrt{x}}\,dx, \qquad (1)$$

где $c > 0$ — некоторое, пока произвольное число. Для вычисления первого интеграла, входящего в правую часть равенства (1), воспользуемся разложением

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \ldots$$

Тогда

$$\int\limits_0^{c} \frac{\cos x}{\sqrt{x}}\,dx = \int\limits_0^{c} \frac{dx}{\sqrt{x}} - \int\limits_0^{c} \frac{x^{3/2}}{2!}\,dx + \int\limits_0^{c} \frac{x^{7/2}}{4!}\,dx - \ldots \qquad (2)$$

Каждый из интегралов в правой части уравнения (2) вычисляется по формуле Ньютона–Лейбница. Пусть задано $\varepsilon > 0$ и требуется вычислить интеграл с погрешностью, не превосходящей заданное значение ε. Для этого достаточно ограничиться суммой нескольких первых членов ряда в правой части уравнения (2). Число этих членов зависит от $\varepsilon > 0$ и от c. Не стоит брать c большим, например $c = 10$. Тогда даже при не очень малом ε, например при $\varepsilon = 10^{-3}$, потребуется много членов ряда. Не стоит брать $c \ll 1$, так как в этом случае интеграл (2) будет вычисляться экономно, но зато подынтегральная функция в интеграле $\int\limits_c^{10} \frac{\cos x}{\sqrt{x}}\,dx$ будет иметь большие по абсолютной величине производные. Это потребует больших значений n в формуле трапеций (или Симпсона) для достижения заданной точности $\varepsilon = 10^{-3}$ при вычислении этого интеграла.

Для преодоления трудности, вызываемой особенностью, иногда удается воспользоваться заменой переменной. Например, $\int\limits_0^{1} \frac{\cos x}{\sqrt{x}}\,dx$ в результате замены $x = t^2$ переходит в определенный интеграл без особенностей

$$\int\limits_0^{1} \frac{\cos t^2}{t} \cdot 2t\,dt = 2\int\limits_0^{1} \cos t^2 dt.$$

Залог успешного решения задачи — в разумном распределении вычисления между аналитическими преобразованиями и приемами, которые должен выполнить исследователь, и рутинными вычислениями, которые должна выполнить машина. Это вообще характерно для решения любых задач с помощью быстродействующей ЭВМ.

Задача

Предложить алгоритмы для вычисления следующих несобственных интегралов:

$$\int_0^1 \frac{\sin x}{x^{3/2}}\,dx, \quad \int_0^{10} \frac{\ln(1+x)}{x^{3/2}}\,dx, \quad \int_0^{\infty} e^{-x^2}\,dx, \quad \int_0^{\pi/2} \frac{\cos x}{\sqrt{\pi/2 - x}}\,dx.$$

§ 3. Кратные интегралы

Вычисление кратных интегралов вида

$$I^{(m)} = I^{(m)}(f, D) = \int\limits_{D}\!\!...\!\int f(x_1, ..., x_m)\,dx_1 ... dx_m \tag{1}$$

по некоторой области D пространства $x_1, ..., x_m$ можно свести к использованию квадратурных формул для вычисления определенных интегралов вида $\int\limits_a^b f(x)\,dx$. Однако сложность соответствующих алгоритмов и число арифметических операций для получения ответа с заданной точностью $\varepsilon > 0$, вообще говоря, быстро возрастают с ростом размерности n, даже если область интегрирования — единичный куб $D = \{|x_i| < 1,\ i = 1, 2, ..., n\}$.

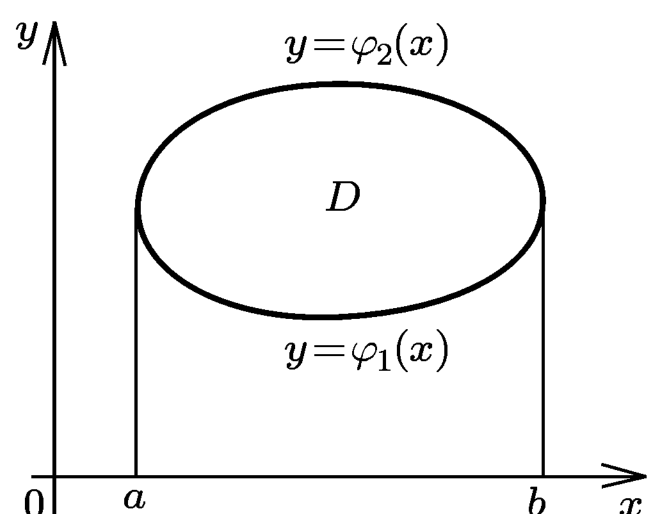


Рис. 7

Большое дополнительное усложнение возникает в случае области интегрирования D общего вида.

1. Переход от кратного интеграла к повторному и использование квадратурных формул. Для выяснения сути дела предположим, что $m = 2$, и рассмотрим интеграл вида

$$I^{(2)} = \iint\limits_{D} f(x, y)\,dxdy \tag{2}$$

по области D (рис. 7), заключенной между кривыми

$$y = \varphi_1(x), \quad y = \varphi_2(x). \tag{3}$$

Очевидно, что

$$I^{(2)} = \iint\limits_{D} f(x, y)\,dxdy = \int\limits_a^b F(x)\,dx, \tag{4}$$

где

$$F(x) = \int\limits_{\varphi_1(x)}^{\varphi_2(x)} f(x, y)\,dy, \quad a \leqslant x \leqslant b. \tag{5}$$

Интеграл (5) можно приближенно заменить квадратурной формулой. Возьмем для определенности формулу трапеций вида

$$I^{(2)} = \int\limits_a^b F(x)\,dx \approx \frac{b-a}{n}\Big(\frac{F(x_0)}{2} + F(x_1) + ... + F(x_{n-1}) + \frac{F(x_n)}{2}\Big),$$

$$x_k = a + k\,\frac{b-a}{n}, \quad k = 0, 1, ..., n. \tag{6}$$

Выражения

$$F(x_j) = \int\limits_{\varphi_1(x_j)}^{\varphi_2(x_j)} f(x_j, y)\,dy,$$

входящие в формулу (6), в свою очередь можно вычислять по квадратурной формуле трапеций:

$$F(x_j) \approx \frac{\varphi_2(x_j) - \varphi_1(x_j)}{n}\times$$

$$\times\Big(\frac{f(x_j, y_0)}{2} + f(x_j, y_1) + ... + f(x_j, y_{n-1}) + \frac{f(x_j, y_n)}{2}\Big), \tag{7}$$

$$y_k = \varphi_1(x_j) + \frac{\varphi_2(x_j) - \varphi_1(x_j)}{n}\,k, \quad k = 0, 1, ..., n.$$

Погрешность $\varepsilon = \varepsilon(n)$ при приближенном вычислении интеграла $I^{(2)}$ по формулам (6), (7) уменьшается с ростом n, а число арифметических операций $\nu(n)$ возрастает. Очевидно, что $\nu(n) = O(n^2)$. Если функция $f(x, y)$ имеет ограниченные производные второго порядка, а функции $\varphi_1(x)$, $\varphi_2(x)$ достаточно гладкие, то в силу свойств формулы трапеций $\varepsilon(n) = O(n^{-2})$.

Прием перехода от кратного интеграла к повторному, а затем вычисления последнего с помощью многократного использования квадратурных формул может быть перенесен и на общий случай кратных интегралов $I^{(m)}$ $(m > 2)$, если область D задана достаточно удобно. При этом число $\nu(n)$ арифметических операций при использовании квадратурных формул с заданным n есть

$$\nu(n) = O(n^m).$$

Если $f(x_1, x_2, ..., x_m)$ имеет ограниченные производные второго порядка, граница Γ области D достаточно гладкая и в качестве квадратурных формул многократно используется формула трапеций, погрешность ε при вычислении кратного интеграла есть $\varepsilon = \varepsilon(n) = O(n^{-2})$.

При заданном ε выбор n производится обычно путем экспериментальных расчетов последовательно с некоторым $n = n_0, n = 2n_0, ...$ до тех пор, пока результат перестает изменяться в пределах заданной точности. Существуют приемы ускорения расчетов.

2. Эффективные методы для вычисления кратных интегралов при больших значениях размерностей m. Приведем простейший пример, где упрощение достигается за счет учета специфики задачи. Пусть требуется вычислить интеграл $\iint\limits_D f(x,y)\,dxdy$, причем известно, что D — круг: $x^2+y^2 \leqslant R^2$, а $f(x,y)$ в действительности зависит не от x и y в отдельности, а от комбинации $x^2+y^2=r^2$, так что $f(x,y) \equiv f(r)$. Тогда

$$I^{(2)} = \iint\limits_D f(x,y)\,dxdy = 2\pi \int\limits_0^R f(r)r\,dr,$$

и дело свелось к вычислению обычного определенного интеграла.

О вычислении кратных интегралов см. [10, 22].

3. Понятие о методе Монте-Карло. Указанный выше прием редукции к обычным определенным интегралам окажется затруднительным даже в случае $m=2$, если область D ограничена достаточно сложной кривой.

Будем считать, что область D лежит внутри квадрата $(0<x<1$, $0<y<1)$. Доопределим функцию $f(x,y)$ вне D нулем. Пусть имеется датчик случайных чисел, который выдает числа, равномерно распределенные на отрезке $[0,1]$ с вероятностью, равной единице. С помощью этого датчика строим последовательность точек (пар чисел) $P_k=(x_k,y_k)$ $(k=1,2,...)$ и составляем числовую последовательность

$$S_k = \frac{1}{k}\sum_{j=1}^{k} f(x_j,y_j).$$

Легко представить, что при произвольном $\varepsilon>0$ вероятность того, что погрешность $I^{(2)}-S_k$ меньше ε:

$$|I^{(2)}-S_k|<\varepsilon,$$

с ростом числа k испытаний стремится к единице. Мы не приводим здесь существующих точных теорем.

Смысл метода Монте-Карло особенно нагляден, если $f(x,y)=1$. Тогда $I^{(2)}$ есть площадь области D, а S_k — отношение числа тех из точек $P_1, P_2, ..., P_k$, которые попали в область D, к числу k всех этих точек. Очевидно, что это отношение приблизительно равно площади области D, поскольку площадь всего квадрата, куда попадают точки, равна единице.

Привлекательность метода Монте-Карло — в логической простоте соответствующего алгоритма, который не усложняется с ростом размерности m и с усложнением формы области D. Однако в случае простых областей и узких классов функций он проигрывает методам, учитывающим эту специфику.

Метод Монте-Карло имеет и многие другие приложения [23].

ЧАСТЬ II

СИСТЕМЫ СКАЛЯРНЫХ УРАВНЕНИЙ

Числа называют еще скалярами, а поэтому уравнения и системы уравнений относительно одного или нескольких неизвестных чисел называют скалярными. Существуют также функциональные уравнения, в частности, дифференциальные уравнения, в которых неизвестными являются не числа, а функции, но в этой части книги мы ими заниматься не будем.

Среди систем скалярных уравнений можно выделить важный класс систем линейных алгебраических уравнений, которому уделено много внимания (гл. 4, 5).

В гл. 6 рассматриваются нелинейные скалярные уравнения.

ГЛАВА 4

СИСТЕМЫ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ. МЕТОДЫ ОТЫСКАНИЯ ТОЧНОГО РЕШЕНИЯ

Системы линейных алгебраических уравнений (СЛАУ) высокого порядка возникают во многих приложениях. Рассмотрим координатную форму записи СЛАУ порядка n:

$$\begin{gathered} a_{11}x_1 + ... + a_{1n}x_n = f_1, \\ \dotfill \\ a_{n1}x_1 + ... + a_{nn}x_n = f_n, \end{gathered}$$

где a_{ij}, f_i $(i, j = 1, 2, ..., n)$ — заданные числа, x_j $(j = 1, 2, ..., n)$ — искомые числа. Более общая и часто более удобная форма задания СЛАУ — операторная.

Мы обсудим формы записи совместных СЛАУ (§ 1), определим понятие обусловленности оператора и соответствующей СЛАУ (§ 2), рассмотрим некоторые точные (§ 4, 6, 7) и итерационные (§ 4) методы вычисления решений СЛАУ, указав классы СЛАУ, для которых один алгоритм вычисления решений предпочтительнее другого. В § 5 устанавливается связь между задачей на минимум квадратичной функции и СЛАУ.

Отметим сразу же, что для вычисления решения СЛАУ известные формулы Крамера при сколько-нибудь большом n $(n > 5)$ не применяются из-за большого числа арифметических операций и вычислительной неустойчивости.

§ 1. Формы записи совместных СЛАУ

Линейным уравнением относительно неизвестных $z, u, ..., w$ называется уравнение вида

$$\alpha z + \beta u + ... + \gamma w = t, \qquad (1)$$

где $\alpha, \beta, ..., \gamma$, а также t — какие-нибудь заданные числа.

1. Каноническая форма записи СЛАУ. Пусть задана система n линейных уравнений относительно того же числа n неизвестных. Эту систему, очевидно, можно записать в следующей (канонической) форме:

$$\begin{aligned} a_{11}x_1 + ... + a_{1n}x_n &= f_1, \\ \dots\dots\dots\dots&\dots\dots \\ a_{n1}x_1 + ... + a_{nn}x_n &= f_n. \end{aligned} \qquad (2)$$

Для этого надо занумеровать заданные уравнения вида (1) и неизвестные номерами $1, 2, ..., n$, а затем неизвестные переобозначить через x_j, коэффициент при неизвестном x_j в уравнении, получившем номер i, — через a_{ij}, а правую часть — через f_i $(i, j = 1, 2, ..., n)$. Из линейной алгебры известно, что система (2) имеет решение (совместна) при любых правых частях в том и только том случае, если соответствующая однородная система имеет только тривиальное решение $x_1 = x_2 = ... = x_n = 0$. Необходимым и достаточным условием этого является условие, чтобы матрица A коэффициентов системы

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \qquad (3)$$

имела определитель, отличный от нуля: $\det A \neq 0$. Условие совместности системы (2) при любых правых частях $\det A \neq 0$ одновременно обеспечивает и единственность решения.

С помощью матрицы A систему (2) можно переписать в виде

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} f_1 \\ \dots \\ f_n \end{bmatrix}, \qquad (4)$$

или, короче,

$$Ax = f. \qquad (5)$$

2. Операторная форма записи. Пусть заданы линейное пространство R^n размерности n, линейный оператор $A : R^n \to R^n$ и некоторый элемент $f \in R^n$. Рассмотрим уравнение

$$Ax = f \tag{6}$$

относительно $x \in R^n$. Известно, что это уравнение имеет решение при произвольном $f = R^n$ в том и только том случае, если соответствующее однородное уравнение $Ax = 0$ имеет только тривиальное решение $x = 0 \in R^n$.

Пусть R^n состоит из элементов вида

$$z = \begin{bmatrix} z_1 \\ \dots \\ z_n \end{bmatrix}, \tag{7}$$

где $z_1, \dots, z_n$ — всевозможные упорядоченные системы чисел. В случае задания элементов пространства R^n в форме (7) будем говорить, что *элементы заданы своими координатами.*

Из линейной алгебры известно, что в этом случае каждому линейному оператору A соответствует матрица вида (3) такая, что заданному $x \in R^n$ сопоставляется $y = Ax$ по формулам

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix}. \tag{8}$$

Уравнение (6) в этом случае совпадает с уравнением (4), или с канонической записью (2). Оно является, таким образом, лишь операторной интерпретацией системы (2). Однако элементы пространства R^n не обязательно задаются как упорядоченные системы чисел (7), а оператор A не обязательно исходно задается в форме (8).

Таким образом, операторная форма задания (6) системы линейных уравнений может отличаться от канонической формы (2).

Приведем пример прикладной задачи, которая естественно приводит к системе линейных уравнений высокого порядка, заданной в операторной форме. Этот пример будет нами многократно использоваться в книге.

3. Пример. Разностный аналог задачи Дирихле для уравнения Пуассона. Пусть в квадратной области $D = \{0 < x,\, y < 1\}$ требуется численно решить следующую задачу:

$$-\Delta u = -\left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) = f(x, y), \quad (x, y) \in D, \tag{9}$$

$$u\big|_{\partial D} = 0,$$

где ∂D — граница области D. Для численного решения этой задачи зададим натуральное M, введем $h = M^{-1}$ и построим сетку с узлами

$$(x_{m_1},\, y_{m_2}) = (m_1 h,\, m_2 h), \quad m_1, m_2 = 0, 1, \dots, M. \tag{10}$$

Будем искать таблицу приближенных значений $u_{m_1 m_2}$ решения $u(x, y)$ задачи (9) в узлах сетки (10) (рис. 8). Для этого в каждом узле сетки $(x_{m_1}, y_{m_2}) = (m_1 h, m_2 h)$, лежащем внутри D, заменим

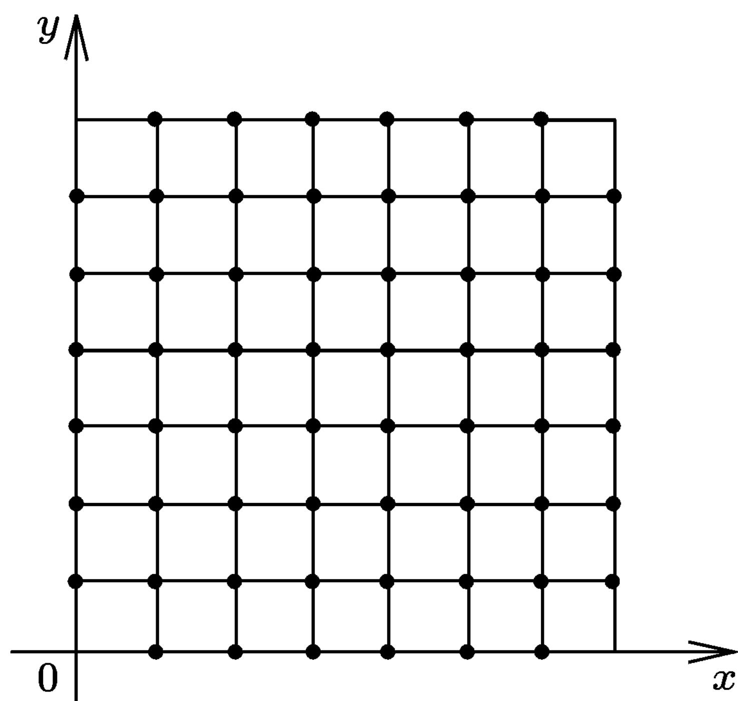


Рис. 8

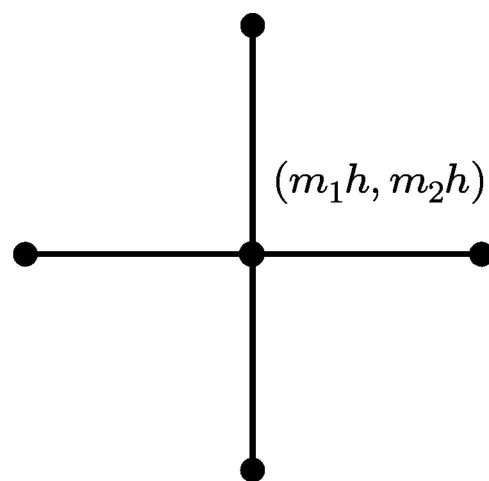


Рис. 9

дифференциальное уравнение (3) разностным, заменяя производные $\partial^2 u/\partial x^2\big|_{(x_{m_1}, y_{m_2})}$, $\partial^2 u/\partial y^2\big|_{(x_{m_1}, y_{m_2})}$ разностными отношениями

$$-\Delta^{(h)} u\big|_{m_1 m_2} = -\Big(\frac{u_{m_1+1, m_2} - 2u_{m_1 m_2} + u_{m_1-1, m_2}}{h^2} + \frac{u_{m_1, m_2+1} - 2u_{m_1 m_2} + u_{m_1, m_2-1}}{h^2}\Big) = f_{m_1 m_2}, \quad m_1, m_2 = 1, ..., M-1, \tag{11}$$

где $f_{m_1 m_2} = f(x_{m_1}, y_{m_2})$.

Это уравнение, составленное в точке $(m_1 h, m_2 h)$, связывает значение искомой сеточной функции в пяти точках (рис. 9).

Если точка сетки $(m_1 h, m_2 h)$ лежит на расстоянии шага сетки h от границы квадрата D, то в уравнение (11) входят значения в одной или двух точках границы. Значения $u_{m_1 \pm 1, m_2 \pm 1}$ в таких точках положим равными нулю. Таким образом, система (11) есть система $n = (M-1)^2$ линейных уравнений относительно такого же числа неизвестных $u_{m_1 m_2}$ $(m_1, m_2 = 1, 2, ..., M-1)$. Для получения решения задачи Дирихле с большой точностью нужно взять достаточно мелкую сетку, или большое число M. Система (11) будет тогда системой высокого порядка.

Придадим системе (11) смысл операторной формы записи (6). Для этого введем пространство $U^{(h)} = R^n$ $(n = (M-1)^2)$, состоящее из всевозможных вещественных функций $z_{m_1 m_2}$, заданных на сетке $(m_1 h, m_2 h)$ $(m_1, m_2 = 1, 2, ..., M-1)$. Введем оператор $A = -\Delta^{(h)}$, который действует из $U^{(h)}$ в $U^{(h)}$, сопоставляя заданной функции

$u^{(h)} \in U^{(h)}$ некоторую $v^{(h)} \in U^{(h)}$ по формулам

$$v^{(h)}\Big|_{m_1 m_2} \equiv v_{m_1 m_2} = -\Delta^{(h)} u^{(h)}\Big|_{m_1 m_2} =$$
$$= -\frac{u_{m_1+1,\,m_2} - 2u_{m_1,\,m_2} + u_{m_1-1,\,m_2}}{h^2} - \frac{u_{m_1,\,m_2+1} - 2u_{m_1,\,m_2} + u_{m_1,\,m_2-1}}{h^2}. \qquad (12)$$

Система линейных уравнений (11) приобретает тогда вид (6):

$$-\Delta^{(h)} u^{(h)} = f^{(h)}, \qquad (13)$$

где $f^{(h)} = \{f_{m_1 m_2}\} \in U^{(h)} = F^{(h)}$ — заданная функция, а $u^{(h)} \in U^{(h)}$ — искомая функция.

Замечание 1. Систему (11), как всякую систему n линейных уравнений относительно n неизвестных, можно было бы записать в каноническом виде (2). Однако при любой нумерации уравнений и неизвестных матрица возникающей системы (2) уступала бы по простоте и обозримости заданию системы в исходной форме (11) и вполне адекватной ей операторной форме (13).

Замечание 2. Абстрактное уравнение в операторной форме (6) всегда можно привести к каноническому виду (2). Для этого надо выбрать какой-либо базис $e_1, e_2, ..., e_n$ в пространстве R^n, записать элементы x, f из R^n в виде

$$x = x_1 e_1 + ... + x_n e_n, \quad f = f_1 e_1 + ... + f_n e_n$$

и отождествить эти элементы с наборами их координат, так что

$$x = \begin{bmatrix} x_1 \\ ... \\ x_n \end{bmatrix}, \quad f = \begin{bmatrix} f_1 \\ ... \\ f_n \end{bmatrix}.$$

Оператор A, входящий в (6), запишется соответствующей этому оператору и сделанному выбору базиса матрицей (3) по формулам вида (8). Уравнение (6) приобретет канонический вид (4) или (2). Однако преобразование (6) к виду (2) с целью вычисления решения может оказаться трудоемким и нецелесообразным.

Задача

Отображение A: $R^2 \to R^2$ элементов $x = (x_1, x_2)$ пространства R^2 в себя задано следующим образом. Рассматривается следующая задача Коши для системы дифференциальных уравнений:

$$dz_1/dt = z_1, \quad dz_2/dt = z_2, \quad 0 \leqslant t \leqslant 1,$$
$$z_1\Big|_{t=0} = x_1, \quad z_2\Big|_{t=0} = x_2.$$

Принимаем за $y = Ax$ элемент $y = (y_1, y_2)$: $y_1 = z_1(t)\Big|_{t=1}$, $y_2 = z_2(t)\Big|_{t=1}$.

Доказать, что оператор A линейный.

а) Выписать матрицу оператора A, если за базис в пространстве R^2 принять

$$e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

б) То же, что и в п. а), но в базисе

$$e_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

в) Вычислить $x = A^{-1}y$, $y = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$.

§ 2. Нормы

Решение системы линейных уравнений порядка n, а также погрешность приближенного решения можно трактовать как элементы линейного пространства R^n. Для количественных суждений о погрешности нужно ввести понятие длины вектора (элемента) из пространства R^n. Вместо термина “длина вектора” в линейной алгебре употребляется термин “норма вектора”. Напомним определения линейного нормированного пространства R^n и нормы линейного оператора A: $R^n \to R^n$.

1. Нормированные пространства. Будем говорить, что линейное пространство R^n *нормировано*, если каждому элементу $x \in R^n$ сопоставлено неотрицательное число $||x||$, причем выполнены следующие условия (аксиомы).

1°. $||x|| > 0$, если $x \neq 0$.

2°. Для любого числа λ и любого $x \in R^n$

$$||\lambda x|| = |\lambda| \, ||x||.$$

3°. Для любых $x \in R^n$, $y \in R^n$ выполнено неравенство треугольника

$$||x + y|| \leqslant ||x|| + ||y||.$$

Приведем примеры норм. Пусть R^n состоит из элементов вида $x = (x_1, x_2, ..., x_n)$, где x_j — числа. Можно показать, что функции $||x||_1$, $||x||_2$, определенные равенствами

$$||x||_1 = \max_j |x_j|, \tag{1}$$

$$||x||_2 = \sum_j |x_j|, \tag{2}$$

удовлетворяют аксиомам 1°–3°. Они называются *первой и второй нормами*.

Введем в R^n *скалярное умножение* (x, y), положив

$$(x, y) = x_1 y_1 + ... + x_n y_n \tag{3}$$

в случае вещественного пространства или

$$(x, y) = x_1 \overline{y}_1 + ... + x_n \overline{y}_n \tag{4}$$

в случае комплексного пространства.

Напомним, что вещественное линейное пространство со скалярным умножением называется *евклидовым*, а комплексное — *унитарным.*

Можно проверить, что функция

$$\|x\|_3 = (x, x)^{1/2} \tag{5}$$

является нормой (удовлетворяет условиям 1°–3°). В случае вещественного пространства эту норму называют *евклидовой*, а в случае комплексного — *эрмитовой.*

Мы привели примеры норм в случае, если элементы пространства R^n записаны как векторы $x = \begin{bmatrix} x_1 \\ ... \\ x_n \end{bmatrix}$, т. е. заданы своими координатами. Возможно задание норм и без использования координатной записи элементов. Так, в пространстве $U^{(h)} = R^n$ $(n = (M-1)^2)$ функций $u_{m_1 m_2}$ $(m_1, m_2 = 1, 2, ..., M-1)$, которое мы ввели и использовали при рассмотрении разностного аналога уравнения Пуассона в § 1, можно ввести нормы, положив

$$\|u^{(h)}\| = \max_{m_1, m_2} |u_{m_1 m_2}|, \tag{6}$$

$$\|u^{(h)}\| = \sum_{m_1, m_2} |u_{m_1 m_2}|. \tag{7}$$

Можно проверить, что в вещественном пространстве $U^{(h)}$ можно ввести скалярное умножение $(u^{(h)}, v^{(h)})$, положив

$$(u^{(h)}, v^{(h)}) = h^2 \sum_{m_1, m_2} u_{m_1 m_2} v_{m_1 m_2}, \tag{8}$$

а затем определить соответствующую евклидову норму, положив

$$\|u^{(h)}\| = (u^{(h)}, u^{(h)})^{1/2}. \tag{9}$$

Вообще, если в вещественном или комплексном линейном пространстве R^n введено какое-либо скалярное умножение (x, y), то возникает соответствующая евклидова или эрмитова норма

$$\|x\| = (x, x)^{1/2}. \tag{10}$$

Можно описать все скалярные умножения и соответствующие евклидовы (эрмитовы) нормы. Для этого фиксируем какое-либо одно скалярное умножение (x, y).

Напомним, что оператором B^*, *сопряженным* (в смысле выбранного скалярного умножения) какому-либо заданному линейному оператору $B\colon R^n \to R^n$, называется такой линейный оператор $B^*\colon R^n \to R^n$, для которого выполнено тождество

$$(Bx, y) \underset{x,\,y}{=} (x, B^*y).$$

Известно, что для всякого оператора B существует один и только один сопряженный ему оператор B^*.

Оператор B называют *самосопряженным*, если $B^* = B$.

Оператор $B\colon R^n \to R^n$ называется *положительно определенным* $(B > 0)$, если $(Bx, x) > 0$ для всех $x \neq 0$. Известно, что если $B + B^* > 0$, то выражение

$$[x, y]_B \equiv (Bx, y) \tag{11}$$

удовлетворяет аксиомам скалярного умножения, причем каждое скалярное умножение в R^n может быть задано формулой (11) при подходящем подборе $B = B^* > 0$. В соответствии с этим каждый положительно определенный и самосопряженный оператор $B = B^* > 0$ порождает евклидову (эрмитову) норму

$$\|x\|_B = ([x, x]_B)^{1/2}. \tag{12}$$

В частности, евклидову норму (10), порожденную первоначально выбранным скалярным умножением, можно записать в форме (12), положив $B = E$, E — тождественный оператор:

$$\|x\|_E = ([x, y]_E)^{1/2} = (x, x)^{1/2}.$$

2. Норма линейного оператора. Введем *норму* $\|A\|$ *оператора* $A\colon R^n \to R^n$, *согласованную с нормой*, выбранной в пространстве R^n, положив

$$\|A\| = \max_{x \in R^n} \frac{\|Ax\|}{\|x\|}, \tag{13}$$

где $\|x\|$, $\|Ax\|$ — нормы элементов x, Ax из нормированного пространства R^n.

Таким образом, $\|A\|$ — это коэффициент растяжения того вектора $x' \in R^n$, который растягивается не слабее любого другого $x \in R^n$.

Условимся каждую матрицу

$$A = \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{bmatrix} \tag{14}$$

отождествлять с оператором $A\colon R^n \to R^n$, который действует в пространстве R^n, состоящем из элементов x вида $x = \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix}$, где x_j — числа, и сопоставляет заданному x некоторый $y = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$ по формулам

$$y_j = \sum_j a_{ij} x_j, \quad i, j = 1, 2, ..., n. \tag{15}$$

Благодаря этому условию приобретает смысл понятие нормы матрицы A как нормы оператора, задаваемого формулами (15).

Теорема 1. *Для норм матрицы $A = (a_{ij})$, согласованных с нормами $||x||_1 = \max |x_j|$, $||x||_2 = \sum_j |x_j|$, в пространстве R^n векторов $x = \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix}$ справедливы формулы*

$$||A||_1 = \max_i \sum_j |a_{ij}|, \tag{16}$$

$$||A||_2 = \max_j \sum_i |a_{ij}|. \tag{17}$$

Доказательство предоставляем читателю.

Теорема 2. *Пусть R^n — евклидово пространство, скалярное умножение в котором будем обозначать (x, y), и пусть $||x|| = (x, x)^{1/2}$ — соответствующая евклидова норма. Пусть A — самосопряженный оператор $(Ax, y) \equiv (x, Ay)$. Тогда*

$$||A|| = \max \frac{||Ax||}{||x||} = \max_j |\lambda_j|, \tag{18}$$

где λ_j $(j = 1, 2, ..., n)$ — собственные числа оператора A.

Доказательство. Из линейной алгебры известно, что в случае $A = A^*$ существует ортонормированный базис

$$e_1, e_2, ..., e_n, \tag{19}$$

состоящий из собственных векторов оператора

$$Ae_j = \lambda_j e_j, \quad j = 1, 2, ..., n.$$

Запишем произвольный $x \in R^n$, а также Ax в виде линейных комбинаций векторов базиса

$$x = x_1 e_1 + x_2 e_2 + ... + x_n e_n, \tag{20}$$

$$Ax = \lambda_1 x_1 e_1 + \lambda_2 x_2 e_2 + ... + \lambda_n x_n e_n. \tag{21}$$

В силу ортонормированности базиса

$$\|x\| = (x_1^2 + x_2^2 + \ldots + x_n^2)^{1/2},$$
$$\|Ax\| = [(\lambda_1 x_1)^2 + (\lambda_2 x_2)^2 + \ldots + (\lambda_n x_n)^2]^{1/2}.$$

Очевидно,

$$\|Ax\| \leqslant \max_j |\lambda_j|(x_1^2 + x_2^2 + \ldots + x_n^2)^{1/2} = \max|\lambda_j|\|x\|.$$

Поэтому для любого $x \in R^n$

$$\|Ax\|/\|x\| \leqslant \max|\lambda_j|,$$

а для $x = e_k$, $\max|\lambda_j| = |\lambda_k|$, выполнено точное равенство

$$\|Ae_k\|/\|e_k\| = |\lambda_k| = \max_j |\lambda_j|.$$

Следовательно, справедливо (18). □

Задачи

1. В пространстве R^2, состоящем из элементов $x = (x_1, x_2)$, которые будем изображать точками плоскости Ox_1x_2, нарисовать совокупность точек, для которых $\|x\| = 1$ (т. е. единичную окружность), в случае, если норма понимается в каждом из следующих трех смыслов:

$$\|x\| = \|x\|_1 = \max|x_j|,$$
$$\|x\| = \|x\|_2 = |x_1| + |x_2|,$$
$$\|x\| = \|x\|_3 = (x_1^2 + x_2^2)^{1/2}.$$

2. Пусть $A\colon\ R^n \to R^n$ — произвольный линейный оператор, а λ_j — какое-нибудь его собственное число.

Доказать, что тогда при произвольном выборе нормы в R^n соответствующая норма $\|A\|$ оператора A удовлетворяет неравенству $\|A\| \geqslant |\lambda_j|$.

3. Привести пример линейного оператора $A\colon\ R^2 \to R^2$, заданного матрицей с собственными числами $\lambda_1 = \lambda_2 = 1$ и такого, что $\|A\|_k > 1000$ $(k = 1, 2, 3)$.

4. Пусть R^n — евклидово пространство, A, B — произвольные линейные операторы.

Доказать, что тогда $\|AB\| \leqslant \|A\|\,\|B\|$.

5. Пусть R^n — евклидово пространство, A — произвольный оператор, B — ортогональный линейный оператор, т. е. $(Bx, Bx) \underset{x}{\equiv} (x, x)$.

Доказать, что тогда $\|AB\| = \|BA\| = \|A\|$.

6. Пусть R^n — евклидово пространство и $A = A^*$ — самосопряженный линейный оператор $A\colon\ R^n \to R^n$. Тогда $\|A^2\| = \|A\|^2$.

Привести пример, показывающий, что для $A \neq A^*$ возможно $\|A^2\| < \|A\|^2$.

7*. Пусть R^n — евклидово пространство и $A\colon\ R^n \to R^n$ — произвольный линейный оператор. Доказать, что тогда

$$\|A\|^2 = \|AA^*\| = \lambda_{\max}(A^*A),$$

где $\lambda_{\max}(A^*A)$ — наибольшее собственное число оператора $C = A^*A$.

8. Пусть R^2 —двумерное пространство вещественных векторов $x = (x_1, x_2)$ со скалярным умножением $(x, y) = x_1y_1 + x_2y_2$ и нормой $\|x\| = (x, x)^{1/2}$.

Вычислить соответствующую норму следующих матриц:

$$A = \begin{bmatrix} 1 & \sqrt{6} \\ \sqrt{6} & 2 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 1 \\ 5 & 2 \end{bmatrix}.$$

9. Пусть R^2 — двумерное пространство вещественных векторов $x = (x_1, x_2)$ со скалярным произведением $(x, y) = x_1y_1 + x_2y_2$, и пусть $B = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$.

Проверить, что $B = B^* > 0$. Вычислить нормы $\|A\|_B$ матриц A из предыдущей задачи, согласованные с $\|x\|_B^2 = [x, x]_B = (Bx, x)$.

10*. Пусть R^n — евклидово пространство, $\|x\|^2 = (x, x)$ — соответствующая норма в нем и $A\colon R^n \to R^n$ — произвольный линейный оператор.

Доказать, что $\|A\| = \|A^*\|$. Если A — невырожденный оператор, т. е. если A имеет обратный оператор A^{-1}, то $(A^{-1})^* = (A^*)^{-1}$, $\|(A^*)^{-1}\| = \|(A^{-1})^*\|$.

§ 3. Обусловленность СЛАУ

Две на первый взгляд похожие системы линейных уравнений могут обладать существенно различными чувствительностями их решений к погрешностям задания входных данных. Это видно уже для систем второго порядка $Ax = f$:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= f_1, \\ a_{21}x_1 + a_{22}x_2 &= f_2. \end{aligned} \tag{1}$$

Будем считать, что $a_{i1}^2 + a_{i2}^2 = 1$ $(i = 1, 2)$. Решение такой системы

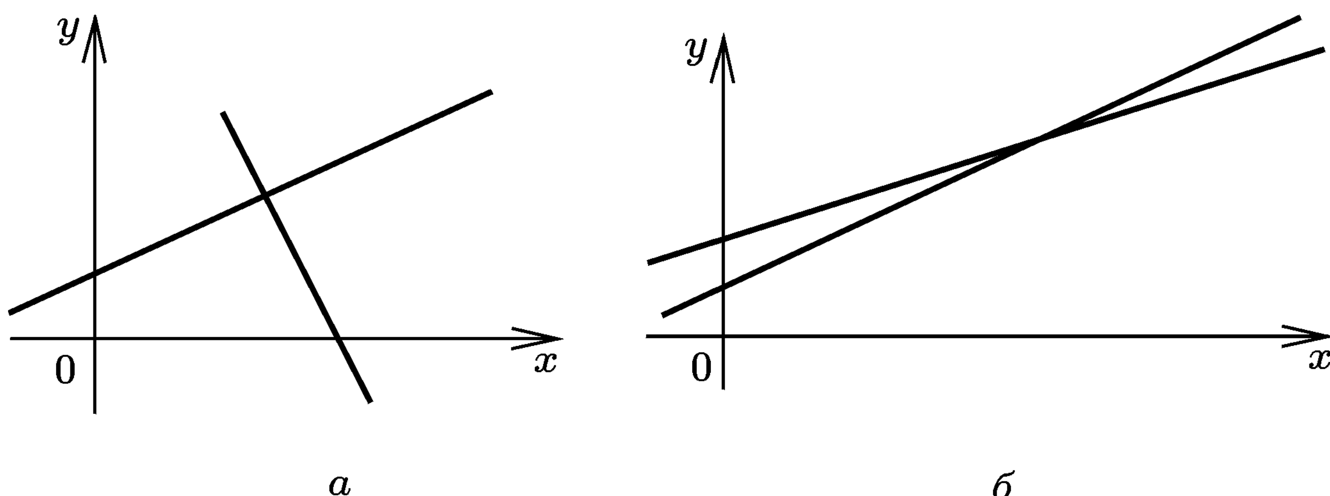


Рис. 10

геометрически интерпретируется как точка пересечения двух прямых на плоскости (x_1, x_2).

Пусть системе (1) соответствует пара прямых: в одном случае, как на рис. 10, *а*, и в другом, как на рис. 10, *б*. Если немного изменить f_1 или коэффициенты a_{11}, a_{22}, то соответствующая прямая

на рис. 10, *а* или 10, *б* сдвинется или повернется. При этом в случае рис. 10, *а* точка пересечения прямых (решение) сдвинется слабо, а в случае рис. 10, *б* — значительно сильнее. Чувствительность решения к возмущению входных данных можно охарактеризовать с помощью так называемого числа $\mu(A)$ обусловленности. В дальнейшем мы увидим, что от числа обусловленности $\mu(A)$ существенно зависит не только чувствительность решения к заданию входных данных, но и число арифметических операций для приближенного вычисления решения (методами последовательных приближений) уравнения $Ax = f$ с заданной точностью.

1. Число обусловленности. *Числом обусловленности линейного оператора* A, действующего в нормированном пространстве R^n, а также *числом обусловленности СЛАУ* $Ax = f$ назовем число

$$\mu(A) = \|A\| \cdot \|A^{-1}\|. \tag{2}$$

Мы условились отождествлять каждую матрицу $A = \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{bmatrix}$ с линейным оператором, действующим в пространстве R^n элементов x вида $x = (x_1, x_2, \ldots, x_n)$, $y = Ax$, где $y = (y_1, y_2, \ldots, y_n)$ вычисляется по формулам

$$y_i = \sum_j a_{ij} x_j.$$

Поэтому приведенное определение числа $\mu(A)$ имеет смысл и для матриц A, так что можно говорить о числе обусловленности матрицы A и о числе обусловленности системы линейных уравнений, заданных не только в операторном, но и в каноническом виде.

Нормы $\|A\|$, $\|A^{-1}\|$ согласованы с нормами, выбранными в R^n. Поэтому число обусловленности $\mu(A)$ также согласовано с выбором нормы в пространстве R^n. Если A — матрица, а для $x \in R^n$ используется норма $\|x\|_1 = \max |x_j|$, то будем писать $\mu_1(A)$; если $\|x\|_2 = \sum |x_j|$, то будем писать $\mu_2(A)$.

Если пространство R^n евклидово со скалярным произведением (x, y) и нормой $\|x\| = (x, x)^{1/2}$, то будем писать $\mu_E(A)$. Если в пространстве R^n на основе исходного фиксированного скалярного умножения (x, y) введены новое скалярное умножение $[x, y]_B = (Bx, y)$ и соответствующая норма $\|x\|_B = ([x, x]_B)^{1/2}$, то будем обозначать соответствующее число обусловленности через $\mu_B(A)$.

Выясним, в чем состоит геометрический смысл числа обусловленности. Для этого рассмотрим совокупность S векторов, норма которых равна единице, т. е. единичную сферу. Среди них отметим по одному вектору $x_{\max}$, $x_{\min}$, для которых верны равенства

$$\|A\, x_{\max}\| = \max_{x \in S} \|Ax\|, \quad \|A\, x_{\min}\| = \min_{x \in S} \|Ax\|.$$

Читатель легко установит, что

$$\|A\| = \|A\, x_{\max}\|, \quad \|A^{-1}\| = 1/\|A\, x_{\min}\|.$$

Отсюда непосредственно следует, что

$$\mu(A) = \max_{x\in S}\|Ax\|/\min_{x\in S}\|Ax\|. \tag{3}$$

Теперь видно, что всегда

$$\mu(A) \geqslant 1. \tag{4}$$

Если A — вырожденный оператор, т. е. A^{-1} не существует, то полагаем $\mu(A) = \infty$.

Геометрический смысл числа $\mu(A)$ особенно нагляден, если используется евклидова норма $\|x\| = (x, x)^{1/2}$, а размерность пространства R^n есть 2, т. е. R^n — плоскость. В этом случае S — единичная окружность: $x_1^2 + x_2^2 = 1$. При линейном преобразовании эта окружность переходит в эллипс. Число $\mu(A)$ в соответствии с (3) есть отношение большой полуоси этого эллипса к его малой полуоси.

Теорема 1. *Пусть оператор $A = A^*$ самосопряжен в смысле скалярного произведения $[x, y]_B$. Тогда*

$$\mu_B(A) = |\lambda_{\max}|/|\lambda_{\min}|, \tag{5}$$

где $\lambda_{\max}$ и $\lambda_{\min}$ — соответственно наибольшее и наименьшее по абсолютной величине собственные числа оператора A.

Доказательство. Пусть $e_1, e_2, ..., e_n$ — ортонормированный в смысле скалярного умножения $[x, y]_B$ базис пространства R^n, состоящий из собственных векторов оператора A, а вещественные числа λ_j ($|\lambda_1| \geqslant |\lambda_2| \geqslant ... |\lambda_n|$) — соответствующие собственные числа, $Ae_j = \lambda_j e_j$. Тогда каждый вектор x можно записать в виде

$$x = x_1e_1 + x_2e_2 + ... + x_ne_n,$$

причем

$$Ax = \lambda_1x_1e_1 + \lambda_2x_2e_2 + ... + \lambda_nx_ne_n,$$
$$\|Ax\| = (|\lambda_1x_1|^2 + |\lambda_2x_2|^2 + ... + |\lambda_nx_n|^2)^{1/2}.$$

Очевидно, что при условии $x \in S$, т. е. при $\|x\| = 1$,

$$\max_{x\in S}\|Ax\|_B = |\lambda_1| = |\lambda_{\max}|, \quad \min_{x\in S}\|Ax\|_B = |\lambda_n| = |\lambda_{\min}|,$$

так что в силу (3) справедливо (5). □

2. Число $\mu(A)$ как характеристика системы $Ax = f$.

Теорема 2. *Пусть правая часть линейного уравнения*

$$Ax = f, \quad x \in R^n, \quad f \in R^n, \tag{6}$$

где A — невырожденный оператор, получила погрешность — приращение Δf. Тогда решение x уравнения получит некоторое приращение Δx, так что

$$A(x + \Delta x) = f + \Delta f. \tag{7}$$

Относительная погрешность $||\Delta x||/||x||$ *решения удовлетворяет неравенству*

$$\frac{||\Delta x||}{||x||} \leqslant \mu(A)\,\frac{||\Delta f||}{||f||}, \tag{8}$$

причем существуют такие f *и* Δf, *при которых в* (8) *достигается строгое равенство.*

Доказательство. Из выражений (7), (6) следует $A(\Delta x) = \Delta f$, $\Delta x = A^{-1}(\Delta f)$. Используем еще $Ax = f$. Тогда

$$\frac{||\Delta x||}{||x||} = \frac{||A^{-1}\Delta f||}{||x||} = \frac{||Ax||}{||x||}\,\frac{||A^{-1}\Delta f||}{||\Delta f||}\,\frac{||\Delta f||}{||Ax||} = \frac{||Ax||}{||x||}\,\frac{||A^{-1}\Delta f||}{||\Delta f||}\,\frac{||\Delta f||}{||f||}. \tag{9}$$

Но

$$\frac{||Ax||}{||x||} \leqslant ||A||, \qquad \frac{||A^{-1}\Delta f||}{||\Delta f||} \leqslant ||A^{-1}||, \tag{10}$$

так что в силу (9), (10) при любых f, Δf справедливо неравенство

$$\frac{||\Delta x||}{||x||} \leqslant ||A||\,||A^{-1}||\,\frac{||\Delta f||}{||f||} = \mu(A)\,\frac{||\Delta f||}{||f||}, \tag{11}$$

т. е. справедливо (8).

Если Δf — тот элемент пространства R^n, для которого

$$||A^{-1}\Delta f||/||\Delta f|| = ||A^{-1}||,$$

а $f = Ax$ — тот элемент $f \in R^n$, для которого

$$||Ax||/||x|| = ||A||,$$

то выражение (9) совпадет с неравенствами (11) и (8), которые при этих f и Δf превращаются в строгие равенства. □

Подчеркнем, что решение уравнения $Ax = f$ не при всех f одинаково чувствительно к возмущению Δf правой части. При заданном фиксированном f может оказаться, что $||Ax||/||x|| \ll ||A||$, так что оценка (9) в этом случае обеспечивает более слабую чувствительность относительной погрешности $||\Delta x||/||x||$ решения к погрешности $||\Delta f||/||f||$, чем неравенство (8).

Для отыскания точного числа обусловленности нужно уметь найти нормы операторов A, A^{-1}. Это обычно очень трудоемко. Если, например, оператор A задан своей матрицей и нас интересует $\mu_1(A)$ или $\mu_2(A)$, нужно найти обратную матрицу A^{-1}, после чего $||A||_1$, $||A^{-1}||_1$ или $||A||_2$, $||A^{-1}||_2$ вычисляются согласно формулам из § 2. Еще труднее находить число обусловленности $\mu_B(A)$ в евклидовой норме, задаваемой каким-либо оператором $B = B^* > 0$. Поэтому часто ограничиваются получением оценок для $\mu(A)$ сверху, используя ту или иную специфику оператора A. В дальнейшем нам встретятся примеры таких оценок для $\mu_B(A)$.

Здесь укажем класс матриц с диагональным преобладанием, для которого удается получить оценку для $\mu_1(A)$, не отыскивая матрицу A^{-1}.

Матрица

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$$

называется *матрицей с диагональным преобладанием* величины $\delta > 0$, если

$$|a_{ii}| \geqslant \sum_{j, j\neq i} |a_{ij}| + \delta, \quad i = 1, 2, \dots, n. \tag{12}$$

Теорема 3. *Пусть A — матрица с диагональным преобладанием величины $\delta > 0$. Тогда существует обратная матрица A^{-1}, причем ее норма, связанная с нормой $\|x\|_1 = \max\limits_j |x_j|$, удовлетворяет оценке*

$$\|A^{-1}\|_1 \leqslant 1/\delta. \tag{13}$$

Доказательство. Зададим произвольно $f = \begin{bmatrix} f_1 \\ \dots \\ f_n \end{bmatrix}$ и предположим, что при этом f система уравнений $Ax = f$ имеет решение $x = \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix}$, причем $\max\limits_i |x_i| = |x_k|$. Выпишем скалярное уравнение с номером k из числа составляющих систему $Ax = f$:

$$a_{k1}x_1 + a_{k2}x_2 + \dots + a_{kn}x_n = f_k.$$

Учтем, что $|x_k| \geqslant |x_i|$ $(i = 1, 2, \dots, n)$, и выпишем оценку

$$|f_k| = \Big|\sum_j a_{kj}x_j\Big| \geqslant |a_{kk}|\,|x_k| - \sum_{j\neq k} |a_{kj}||x_j| \geqslant$$
$$\geqslant |a_{kk}|\,|x_k| - \Big(\sum_{j\neq k} |a_{kj}|\Big)|x_k| = \Big(|a_{kk}| - \sum_{j\neq k} |a_{kj}|\Big)|x_k| \geqslant \delta |x_k|.$$

Отсюда $|x_k| \leqslant \dfrac{1}{\delta}\,|f_k|$. Но $|x_k| = \max\limits_j |x_j| = \|x\|_1$, $|f_k| \leqslant \max\limits_j |f_j| = \|f\|_1$. Поэтому справедливо (13):

$$\|x\|_1 \leqslant \frac{1}{\delta}\,\|f\|_1. \tag{14}$$

В частности, в случае $f = 0 \in R^n$ отсюда следует, что система $Ax = 0$ имеет только тривиальное решение $x = 0$, а значит, система $Ax = f$ имеет одно и только одно решение при любом $f \in R^n$ и существует A^{-1}. $\square$

Оценка (14) означает, что для $x = A^{-1}f$ при любом f имеет место оценка

$$\|A^{-1}f\|_1 \leqslant \frac{1}{\delta}\,\|f\|_1, \qquad \frac{\|A^{-1}f\|_1}{\|f\|_1} \leqslant \frac{1}{\delta},$$

так что

$$\|A^{-1}\|_1 \leqslant \max_f \frac{\|A^{-1}f\|_1}{\|f\|_1} \leqslant \frac{1}{\delta}. \ \square$$

Следствие. *Пусть A — матрица с диагональным преобладанием величины δ. Тогда*

$$\mu_1(A) = \|A\|_1\|A^{-1}\|_1 \leqslant \frac{1}{\delta}\|A\|_1. \tag{15}$$

Задачи

1. Доказать, что числа обусловленности $\mu_1(A)$, $\mu_2(A)$ матрицы A не изменятся, если в матрице A поменять местами строки или столбцы.

2. Доказать, что для матрицы $A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$ и соответствующей транспонированной матрицы $A^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{bmatrix}$ выполнены равенства $\mu_1(A) = \mu_2(A^T)$, $\mu_2(A) = \mu_1(A^T)$.

3. Показать, что число обусловленности оператора A не меняется при умножении этого оператора на произвольное вещественное число k $(k \neq 0)$.

4. Показать, что $\mu_B(A) = 1$ в том и только том случае, если выполнено одно из условий:

а) A — оператор подобия, т. е. $Ax \underset{x}{\equiv} kx$ $(k^2 > 0)$;

б) A — ортогональный оператор, т. е. $[Ax, Ax]_B \equiv [x, x]_B$;

в) A — произведение оператора подобия и ортогонального оператора.

5*. Показать, что $\mu_B(A) \underset{A}{\equiv} \mu_B(A^*_B)$, где A^*_B — оператор, сопряженный оператору A в смысле скалярного умножения $[x, y]_B$.

6. Пусть A — матрица, причем $\det A \neq 0$. Умножим одну из строк на некоторое число k и обозначим результат через A_k.

Показать, что $\mu(A_k) \to \infty$ при $k \to \infty$.

7*. Доказать, что для любого линейного оператора A

$$\mu_B(A^*_B A) = (\mu_B(A))^2,$$

где A^*_B — оператор, сопряженный оператору A в смысле скалярного умножения $[x, y]_B$.

8*. Пусть $A = A^*$, $B = B^* > 0$ в смысле некоторого скалярного умножения (x, y). Пусть для всех $x \in R^n$ выполнены неравенства

$$\gamma_1(Bx, x) \leqslant (Ax, x) \leqslant \gamma_2(Bx, x),$$

$\gamma_1 > 0$, $\gamma_2 > 0$ — некоторые числа. Рассмотрим оператор $C = B^{-1}A$.

Доказать, что справедливо неравенство

$$\mu_B(C) \leqslant \gamma_2/\gamma_1.$$

Примечание. Решение этой задачи мы приведем в гл. 5 в связи с ее приложениями.

§ 4. Методы исключения Гаусса

Опишем простой метод исключения Гаусса и метод Гаусса с выбором главного элемента для точного решения систем линейных уравнений, заданных в каноническом виде

$$\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = f_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = f_n. \end{array} \quad (1)$$

1. Простой метод Гаусса. Из первого уравнения системы (1) выражаем x_1 через остальные. Получим

$$x_1 = a'_{12}x_2 + \ldots + a'_{1n}x_n + f'_1. \quad (2)$$

Подставим это выражение вместо x_1 в остальные $n-1$ уравнений и получим систему $n-1$ уравнений относительно $n-1$ неизвестных $x_2, x_3, \ldots, x_n$. Из первого уравнения этой системы выразим x_2 через остальные:

$$x_2 = a'_{23}x_3 + \ldots + a'_{2n}x_n + f'_2. \quad (3)$$

Действуя аналогично, при $k = 3, 4, \ldots, n$ получаем

$$x_k = a'_{k,k+1}x_{k+1} + \ldots + a'_{kn}x_n + f'_k, \quad (4)$$

причем очевидно, что при $k = n$

$$x_n = f'_n. \quad (5)$$

Равенство (5) дает значение x_n, а затем по формулам (4) последовательно найдем $x_{n-1}, x_{n-2}, \ldots, x_1$. Указанный алгоритм может оказаться нереализуемым из-за деления на нуль или дать грубую ошибку в результате округлений. Объясним это.

Если x_1 входит в первое уравнение системы (1) с нулевым коэффициентом a_{11}, то уже запись (2) невозможна, так как $a'_{12} = -a_{12}/a_{11}$. Необходимость деления на нуль может встретиться на любом шаге процесса. Если деление на нуль не встречалось и формулы (4) при $k = 1, 2, \ldots, n$ получены, то может возникнуть вычислительная неустойчивость при вычислении $x_n, x_{n-1}, \ldots, x_1$. Если, например, $a'_{k,k+1} = 2$ ($k = n-1, n-2, \ldots, 1$), а остальные a'_{ij} обращаются в нуль, то погрешность ε, допущенная при вычислении x_n, возрастет вдвое при вычислении x_{n-1}; еще вдвое возрастет при вычислении x_{n-2} и в 2^{n-1} раз при вычислении x_1. Уже при $n = 10$ погрешность возрастает в тысячу раз.

Нужно иметь в виду и опасность того, что числа f'_k могут быстро возрастать с ростом k. Если бы это было так, то малые относительные погрешности, допущенные при вычислении f'_k по формуле (4), вносили бы большую по абсолютной величине погрешность в x_k.

Сформулируем достаточное условие, гарантирующее вычислительную устойчивость метода Гаусса.

Теорема 1. *Пусть матрица A системы* (1) *является матрицей с диагональным преобладанием* (*величины* $\delta > 0$). *Тогда в алгоритме простого метода Гаусса не встретится деление на нуль, причем выполнены неравенства*

$$\sum_{j=1}^{n-k} |a'_{k,k+j}| < 1, \quad k = 1, 2, ..., n, \tag{6}$$

$$|f'_k| \leqslant \frac{2}{\delta} \max_j |f_j|. \tag{7}$$

Доказательству теоремы предпошлем следующую лемму.

Лемма 1. *Пусть*

$$\begin{aligned} b_{11}t_1 + b_{12}t_2 + ... + b_{1r}t_r &= \varphi_1, \\ \dots\dots\dots\dots\dots\dots \\ b_{r1}t_1 + b_{r2}t_2 + ... + b_{rr}t_r &= \varphi_r \end{aligned} \tag{8}$$

есть некоторая система линейных уравнений с диагональным преобладанием величины δ:

$$|b_{ll}| \geqslant \sum_{j \neq l} |b_{lj}| + \delta. \tag{9}$$

Тогда при получении из первого уравнения системы (8) *выражения для* t_1:

$$t_1 = b'_{11}t_2 + ... + b'_{1r}t_r + \varphi'_1, \tag{10}$$

не встретится деление на нуль; *выполнено неравенство*

$$\sum_{j=2}^{r} |b'_{1j}| < 1. \tag{11}$$

В случае $r > 1$ *система* $(r-1)$*-го порядка относительно* $t_2, t_3, ...$ $..., t_r$, *полученная из* (8) *исключением* t_1 *с помощью формулы* (10), *окажется системой с диагональным преобладанием* (*той же, что и в* (9), *величины* δ).

Доказательство. В силу (9) для $l = 1$

$$|b_{11}| \geqslant |b_{12}| + |b_{13}| + ... + |b_{1r}| + \delta.$$

Отсюда $b_{11} \neq 0$, выражение (10) со значениями

$$b'_{ij} = -b_{1j}/b_{11}, \quad j = 2, 3, ..., r, \quad \widetilde{\varphi}_1 = \varphi_1/b_{11}$$

имеет смысл, а также выполнено (11).

Пусть $r > 1$. Докажем последнее, третье утверждение леммы. Действительно, подставляя (10) в уравнение с номером j ($j > 1$) системы (8), получаем после приведения подобных уравнение

$$(b_{j2} + b_{j1}b'_{12})t_2 + (b_{j3} + b_{j1}b'_{13})t_3 + \ldots + (b_{jr} + b_{jr}b'_{1r})t_r = \widetilde{\varphi}_j, \quad j = 2, 3, \ldots, r. \tag{12}$$

В этой системе $(r-1)$-го порядка уравнением с номером s ($s = 1, 2, \ldots, r-1$) окажется уравнение

$$(b_{s+1,2} + b_{s+1,1}b'_{12})t_2 + (b_{s+1,3} + b_{s+1,1}b'_{13})t_3 + \ldots \\ \ldots + (b_{s+1,r} + b_{s+1,1}b'_{1r})t_r = \widetilde{\varphi}_{s+1}, \quad s = 1, 2, \ldots, r-1. \tag{13}$$

В матрице системы (13) в строке с номером s стоят числа

$$(b_{s+1,2} + b_{s+1,1}b'_{12}), \quad (b_{s+1,3} + b_{s+1,1}b'_{13}), \quad \ldots, \quad (b_{s+1,r} + b_{s+1,1}b'_{1r}),$$

причем на диагонали окажется элемент $b_{s+1,s+1} + b_{s+1,1}b'_{1,s+1}$.

Покажем, что имеет место диагональное преобладание (величины δ), т. е. что справедливо неравенство

$$|b_{s+1,s+1} + b_{s+1,1}b'_{1,s+1}| \geqslant \sum_{\substack{j=2 \\ j\neq s+1}}^{r} |b_{s+1,j} + b_{s+1,1}b'_{1j}| + \delta. \tag{14}$$

Докажем более сильное неравенство

$$|b_{s+1,s+1}| - |b_{s+1,1}b'_{1,s+1}| \geqslant \sum_{\substack{j=2 \\ j\neq s+1}}^{r} [|b_{s+1,j}| + |b_{s+1,1}b'_{1,j}|] + \delta.$$

Оно равносильно неравенству

$$|b_{s+1,s+1}| \geqslant \sum_{\substack{j=2 \\ j\neq s+1}}^{r} |b_{s+1,j}| + |b_{s+1,1}| \sum_{j=2}^{r} |b'_{1j}| + \delta. \tag{15}$$

В силу неравенства (11) неравенство (15) заведомо будет выполнено, если выполнено неравенство, получающееся из (15) заменой $\sum |b'_{1j}|$ числом 1:

$$|b_{s+1,s+1}| \geqslant \sum_{\substack{j=2 \\ j\neq s+1}}^{r} |b_{s+1,j}| + |b_{s+1,1}| + \delta = \sum_{\substack{j=2 \\ j\neq s+1}}^{r} |b_{s+1,j}| + \delta. \tag{16}$$

Но это неравенство справедливо по условию (9), так что неравенство (14) доказано. □

Доказательство теоремы 1. Сначала докажем формулу (4) и неравенство (6). Воспользуемся индукцией по k. В случае $k = 1$, $n > 1$ справедливость формулы (4) и неравенства (6) установлена в

первых двух утверждениях леммы. При этом в силу третьего утверждения леммы система $(n-1)$-го порядка относительно $x_2, x_3, ..., x_n$, полученная из (1) исключением x_1 с помощью формулы (2), оказывается системой с диагональным преобладанием величины δ. Пусть формула (4) и неравенство (6) уже доказаны для $k = 1, 2, ..., s$ $(s < < n)$, причем доказано также, что система порядка $n - s$ относительно $x_{s+1}, x_{s+2}, ..., x_n$, полученная из (1) исключением $x_1, x_2, ..., x_s$ с помощью формулы (4), обладает диагональным преобладанием величины δ. Рассмотрев эту систему в качестве системы (8) и использовав лемму, установим справедливость индукционного предположения для $k = s + 1$. Доказательство (4), (6) по индукции окончено.

Докажем теперь (7). В силу теоремы 3 из § 3 для решения системы (1) справедливо неравенство

$$\max_j |x_j| \leqslant \frac{1}{\delta} \max |f_j|. \tag{17}$$

Отсюда и из (4) при любом $k = 1, 2, ..., n$ следует неравенство

$$|f'_k| = \Big| x_k - \sum_j a'_{kj} x_j \Big| \leqslant |x_k| + \sum_{j=k+1}^{n} |a'_{kj}| |x_j| \leqslant$$
$$\leqslant \max_{1 \leqslant j \leqslant n} |x_j| \Big(1 + \sum_{j=k+1}^{n} a'_{kj} \Big) \leqslant 2 \max_j |x_j| \leqslant \frac{2}{\delta} \max_j |f_j|,$$

которое совпадает с (7). □

Подчеркнем, что условие теоремы достаточно, но не необходимо для применимости метода Гаусса без выбора главного элемента. Если для какой-либо системы (1) алгоритм метода Гаусса без выбора главного элемента удалось применить на компьютере, то найденное при этом решение не будет точным только в силу конечной разрядности компьютера и связанных с этим неизбежных погрешностей округления.

Подставляя найденное приближенное решение системы (1) в левую часть и вычисляя невязки, можно судить о погрешности решения, воспользовавшись оценкой

$$\frac{\|\Delta x\|}{\|x\|} \leqslant \mu(A) \, \frac{\|\Delta f\|}{\|f\|},$$

если только число обусловленности $\mu(A)$ или оценка для него известны.

2. Порядок числа ν арифметических операций для реализации алгоритма Гаусса, как легко увидит читатель, есть

$$\nu = O(n^3). \tag{18}$$

3. Прогонка. Изложенный метод Гаусса особенно эффективен для системы (1) следующего специального вида:

$$
\begin{array}{lcl}
b_1x_1 + c_1x_2 & = & f_1, \\
a_2x_1 + b_2x_2 + c_2x_3 & = & f_2, \\
a_3x_2 + b_3x_3 + \quad c_3x_4 & = & f_3, \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & & \\
a_{n-1}x_{n-2} + b_{n-1}x_{n-1} + c_{n-1}x_n & = & f_{n-1}, \\
a_nx_{n-1} + \quad b_nx_n & = & f_n,
\end{array}
\qquad (19)
$$

матрица которой трехдиагональна. Условия диагонального преобладания в этом случае принимают вид

$$
\begin{aligned}
&|b_1| \geqslant |c_1| + \delta, \\
&|b_k| \geqslant |a_k| + |c_k| + \delta, \quad k = 2, 3, ..., n-1, \\
&|b_n| \geqslant |a_n| + \delta.
\end{aligned}
\qquad (20)
$$

Уравнения (4) принимают вид

$$x_k = A_k x_{k+1} + F_k, \quad k = 1, 2, ..., n-1, \qquad (21)$$

$$x_n = F_n, \qquad (22)$$

где A_k, F_k — некоторые коэффициенты.

Положим по определению $A_n = 0$ и условимся записывать (21), (22) единой формулой

$$x_k = A_k x_{k+1} + F_k, \quad k = 1, 2, ..., n. \qquad (23)$$

Очевидно, что в (23)

$$A_1 = -c_1/b_1, \quad F_1 = f_1/b_1.$$

Допустим, что уже вычислены A_k, F_k при некотором k ($1 \leqslant k \leqslant n-1$). Подставляя выражение $x_k = A_k x_{k+1} + F_k$ в $(k+1)$-е уравнение системы (13), получаем

$$x_{k+1} = -\frac{c_{k+1}}{b_{k+1} + a_{k+1}A_k}\, x_{k+2} + \frac{f_{k+1} - a_{k+1}F_k}{b_{k+1} + a_{k+1}A_k}.$$

Отсюда получаются рекуррентные соотношения

$$
\begin{aligned}
A_{k+1} &= \frac{-c_k}{b_{k+1} + a_{k+1}A_k}, \\
F_{k+1} &= \frac{f_{k+1} - a_{k+1}F_k}{b_{k+1} + a_{k+1}A_k}, \quad k = 1, 2, ..., n-1.
\end{aligned}
$$

Процесс вычисления решения системы (19) распадается на вычисление коэффициентов A_j, F_j ($j = 1, 2, ..., n$), а затем последовательное вычисление по формулам (23) при $k = n, n-1, ..., 1$ значений $x_n, x_{n-1}, ..., x_1$.

Описанный алгоритм метода Гаусса для системы с трехдиагональной матрицей часто называют *прогонкой*. Коэффициенты A_k, F_k называют *прогоночными коэффициентами*, процесс их вычисления — *прямой прогонкой*. Процесс вычисления компонент решения x_n, $x_{n-1}, ..., x_1$ с помощью прогоночных соотношений (23) называют *обратной прогонкой*.

Подсчитаем порядок числа арифметических действий для реализации прогонки. Для вычисления прогоночных коэффициентов требуется $O(n)$ арифметических операций. Для обратной прогонки требуется n-кратное использование формулы (23), т. е. также $O(n)$ арифметических операций. Общее число арифметических операций есть $O(n)$. Очевидно, не существует более экономного по порядку числа арифметических действий алгоритма для решения системы (19), поскольку число n неизвестных также есть $O(n)$.

Прогонкой называют алгоритм метода Гаусса без выбора главного элемента и в том случае, когда матрица A системы (1) содержит ненулевые элементы на m $(3 < m \ll n)$ соседних диагоналях, среди которых есть главная. Если m фиксировано, а n может быть любым, то число арифметических операций также есть $O(n)$.

Системы вида (19) высокого порядка привлекли особое внимание в начале 50-х годов в связи с тем, что они возникли при использовании так называемых неявных разностных схем для уравнения теплопроводности. О таких схемах и их роли будет рассказано в части III.

И.М. Гельфанд и О.В. Локуциевский предложили алгоритм прогонки для численного решения этих систем, установив его полную адекватность задаче. Они построили континуальное замыкание алгоритма прогонки и установили вычислительную устойчивость этого алгоритма (см. И.М. Гельфанд и О.В. Локуциевский, дополнение к [5]).

Указанная работа была одной из первых, где четко поставлен и решен вопрос об устойчивости вычислительного алгоритма. Этот вопрос приобрел особую остроту при вычислениях на компьютере, когда миллионы и миллиарды операций осуществляются в автоматическом режиме без контроля человека.

Теорема о достаточных условиях применимости метода Гаусса, доказанная в этом параграфе, является обобщением результата И.М. Гельфанда и О.В. Локуциевского об устойчивости прогонки на случай систем с заполненной матрицей.

Отметим, что идею переноса ("прогонки") условия $b_1x_1 + c_2x_2 = f_1$, заданного первым уравнением трехдиагональной системы (19), реализуемую при получении прогоночных коэффициентов и соотношений (23), стали использовать и в других методах исключения, которые также называют методами прогонки [20].

4. Метод Гаусса с выбором главного элемента. Как мы видели, простой метод Гаусса может натолкнуться на препятствия: в процессе его реализации может встретиться деление на нуль или произойти потеря точности из-за вычислительной неустойчивости. Приведем модификацию метода исключения Гаусса — метод Гаусса с выбором главного элемента, который в случае $\det A \neq 0$ гарантирует от деления на нуль и повышает вычислительную устойчивость по сравнению с простым методом Гаусса.

Сначала отыскивается самый большой (один из самых больших, если их несколько) по модулю элементов матрицы A. Допустим, это элемент a_{ij}, стоящий коэффициентом при x_j в уравнении с номером i. Очевидно, что a_{ij} отличен от нуля. Осуществляем перенумерацию уравнений и неизвестных так, чтобы уравнение с номером i стало первым и неизвестное x_j стало первым, $x'_1 = x_j$. После этого осуществляем первый шаг описанного выше простого метода Гаусса (без выбора главного элемента). При рассмотрении полученной системы порядка $n - 1$ вновь производим выбор главного элемента, перенумерацию уравнений и неизвестных и т. д. Получаем систему равенств вида (4), которые позволяют вычислить все компоненты решения, начиная с последней. На компьютерах обычно имеются стандартные программы методов Гаусса.

Простой подсчет показывает, что число арифметических действий метода Гаусса с выбором главного элемента также требует $O(n^3)$ арифметических операций.

Отметим еще, что метод Гаусса позволяет находить обратную матрицу A^{-1}. Столбец с номером j матрицы A^{-1} обозначаем $\begin{bmatrix} x_{1j} \\ \dots \\ x_{nj} \end{bmatrix}$.

Из равенства $AA^{-1} = E$ следует, что этот столбец является решением системы

$$A \begin{bmatrix} x_{1j} \\ \dots \\ x_{nj} \end{bmatrix} = \begin{bmatrix} \delta_{1j} \\ \dots \\ \delta_{nj} \end{bmatrix}, \quad j = 1, 2, \dots, n, \tag{24}$$

где $\delta_{kj} = 0$, если $k \neq j$, $\delta_{jj} = 1$.

Системы (24), которых имеется n экземпляров, различаются только правыми частями. Поэтому большая часть вычислений по методу Гаусса для них совпадает. Порядок числа арифметических операций для отыскания A^{-1} при правильной организации расчета остается $O(n^3)$, как для решения одной системы.

5. Замечание о других универсальных методах точного решения. Существует ряд других методов, в которых тем или иным способом исходная система (1) приводится к более простому виду (в методе Гаусса система (1) приводится к треугольному виду (4)), а затем вычисляется решение этой “простой” системы.

Приведение к этому более простому виду можно осуществить с помощью умножения системы $Ax = f$ на подходящую ортогональную матрицу C (т. е. матрицу, для которой $(Cx, Cx) = (x, x)$). Тогда число обусловленности $\mu_3(CA)$ новой матрицы системы $CAx = Cf$ совпадает с числом обусловленности $\mu_3(A)$ (докажите равенство $\mu_0(CA) = \mu_3(A)$), так что вместо $\mu_3(A)$ можно искать $\mu_3(CA)$, что может оказаться проще. Таким образом, основываясь на теореме 1 из § 3, одновременно с вычислением решения можно оценить погрешность, вызванную неточным заданием входных данных f_i.

6. Об алгоритме с гарантированной оценкой погрешности. При расчете на реальном компьютере с заданным числом разрядов наряду с влиянием неточного задания входных данных на каждой арифметической операции вносятся погрешности округления. Влияние этих погрешностей округления на результат зависит не только от разрядности машины, но и от числа обусловленности матрицы системы, и от избранного алгоритма. В [3] построен алгоритм, который учитывает влияние погрешностей округления на данной машине и выдает результат с гарантированной точностью либо в процессе вычислений устанавливает, что данная система обусловлена настолько плохо, что при расчете на машине с заданной разрядностью какая-либо точность не может быть гарантирована.

Задачи

1. Вычислить решение системы

$$10^{-3}x + y = 5, \quad x - y = 6$$

простым методом Гаусса и с выбором главного элемента. Вести вычисления с двумя значащими цифрами. Сравнить и объяснить результаты.

2. Для численного решения краевой задачи

$$\frac{d^2x}{dt^2} - p^2(t)x = f(t), \quad 0 < t < 1,$$

$$x(0) = \varphi, \quad x(1) = \psi,$$

разобьем отрезок $0 \leqslant x \leqslant 1$ на N равных частей и будем искать приближенно таблицу значений решения $x_0, x_1, ..., x_N$ в точках разбиения $t_n = nh$ $(n = 0, 1, ..., N;\ h = N^{-1})$. В точке t_n $(n = 1, 2, ..., N-1)$ заменим производную разностным отношением

$$\left.\frac{d^2x}{dt^2}\right|_{t=nh} \approx \frac{x_{n+1} - 2x_n + x_{n-1}}{h^2}.$$

Получим вместо исходной задачи ее разностный аналог

$$x_0 = \varphi,$$

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{h^2} - p^2(t_n)\,x_n = f(t_n), \quad n = 1, 2, ..., N-1, \tag{25}$$

$$x_N = \psi.$$

а) Показать, что для решения этой системы линейных уравнений можно и целесообразно воспользоваться прогонкой.

б) Составить программу вычисления решения системы (25) на компьютере, если $p(t) = (1 + t), f(t) = e^t$.

§ 5. Связь между задачей на минимум квадратичной функции и СЛАУ

Пусть R^n есть n-мерное евклидово пространство элементов x. Зададим какой-либо линейный оператор A, $f \in R^n$, число c и рассмотрим функцию от x вида

$$F(x) = (Ax, x) - 2(f, x) + c, \tag{1}$$

называемую *квадратичной функцией.*

Заметим, что эта квадратичная функция в силу $(Ax, x) \equiv \equiv (x, A^*x) \equiv (A^*x, x)$ совпадает с квадратичной функцией

$$F(x) = (A^*x, x) - 2(f, x) + c,$$

а значит, и с функцией

$$F(x) = \left(\frac{A + A^*}{2}x, x\right) - 2(f, x) + c.$$

Без ограничения общности будем считать, что в выражении (1) оператор A самосопряженный. В противном случае мы заменили бы его самосопряженным оператором $(A + A^*)/2$.

Предположим, что $A = A^*$, $A > 0$, т. е. $(Ax, x) > 0$ в случае $x \neq 0$. Поставим задачу об отыскании элемента $z \in R^n$, придающего наименьшее значение функции $F(x)$:

$$F(z) = \min_{x \in R^n} F(x). \tag{2}$$

Задача (2) о минимуме квадратичной функции и задача об отыскании решения системы линейных уравнений

$$Ax = f, \quad A = A^* > 0, \tag{3}$$

равносильны. Сформулируем соответствующее утверждение в виде теоремы.

Теорема 1. *Пусть $A = A^* > 0$. Существует один и только один элемент $z \in R^n$, придающий наименьшее значение квадратичной функции* (1). *Этот элемент есть решение уравнения* (3).

Доказательство. В силу положительной определенности оператора A он невырожден, а следовательно, уравнение (3) имеет одно и только одно решение $z \in R^n$.

Покажем, что при любом $\delta \in R^n$ $(\delta \neq 0)$ будет $F(z + \delta) > F(z)$:

$$F(z + \delta) = (A(z + \delta), z + \delta) - 2(f, z + \delta) + c =$$
$$= [(Az, z) - 2(f, z) + c] + 2(Az, \delta) - 2(f, \delta) + (A\delta, \delta) \equiv$$
$$\equiv F(z) + 2(Az - f, \delta) + (A\delta, \delta) = F(z) + (A\delta, \delta) > F(z). \square$$

Установленная равносильность задач (1) и (3) позволяет сводить решение любой из них к решению другой.

Линейные уравнения вида (3) с самосопряженным и положительно обусловленным оператором A представляют собой важный класс СЛАУ по двум причинам.

Во-первых, СЛАУ

$$Cx = \varphi \tag{4}$$

с произвольным невырожденным линейным оператором $C\colon R^n \to R^n$ сводится к СЛАУ вида (3). Достаточно положить $A = C * C$, $f = C * \varphi$.

Во-вторых, многие краевые задачи для эллиптических уравнений являются задачами Лагранжа–Эйлера для некоторых задач о минимуме квадратичных функционалов. Поэтому естественно, что при "правильной" дискретизации этих вариационных задач возникает задача о минимуме квадратичной функции в конечномерном пространстве, которая в силу доказанной теоремы приводит к СЛАУ (3).

Задачи

1. Задана квадратичная функция скалярных аргументов x_1, x_2:

$$F(x_1, x_2) = x_1^2 + 2x_1x_2 + 4x_2^2 - 2x_1 + 3x_2 + 5.$$

Рассмотреть эту функцию как функцию вектора $x = (x_1, x_2)$ евклидова пространства R^2 со скалярным умножением $(x, y) = x_1y_1 + x_2y_2$ и записать ее в форме (1). Проверить, что $A > 0$, и найти решение соответствующей задачи (3).

2*. Записать функцию $F(x_1, x_2)$ из предыдущей задачи в виде

$$F(x) = [Ax, x]_B - 2[f, x]_B + C,$$

где

$$[x, y]_B = (Bx, y), \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad A = A_B^*.$$

§ 6. Метод сопряженных градиентов как метод точного решения СЛАУ

Рассмотрим уравнение

$$Ax = f, \quad A = A^* > 0, \quad f \in R^n, \quad x \in R^n, \tag{1}$$

где R^n — пространство со скалярным произведением (x, y) и $A\colon R^n \to R^n$ — самосопряженный в смысле этого скалярного умножения положительно определенный оператор.

1. Метод сопряженных градиентов вычисления решения. Зададим произвольно $x_0 \in R^n$ и построим последовательность

$$\begin{aligned} x_1 &= (E - \tau_1 A)\, x_0 + \tau_1 f, \\ x_{k+1} &= \alpha_{k+1}(E - \tau_{k+1} A)\, x_k + (1 - \alpha_{k+1})\, x_{k-1} + \alpha_{k+1}\tau_{k+1} f, \end{aligned} \tag{2}$$

где

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Ar_k, r_k)}, \quad r_k = Ax_k - f, \quad k = 0, 1, ..., \\ \alpha_1 = 1, \quad \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})} \frac{1}{\alpha_k}\right)^{-1}, \quad k = 1, 2, ... \tag{3}$$

Оказывается, что существует номер k_0 ($k_0 \leqslant n$) такой, что член x_{k_0} последовательности (2) совпадает с точным решением x СЛАУ (1):

$$x = x_{k_0}, \quad k_0 \leqslant n. \tag{4}$$

Отметим также, что $x_1, x_2, ..., x_k, ...$ являются уточняющимися с ростом номера k последовательными приближениями к решению; при заданном малом $\varepsilon > 0$ погрешность $\|x - x_k\|$ приближения x_k при условиях, которые будут указаны в § 2 гл. 5, может стать меньше ε уже при $k \ll n$.

В настоящее время метод сопряженных градиентов при "умеренном" числе обусловленности $\mu(A)$ и больших n на практике используется обычно именно как метод последовательных приближений.

Вычисление точного решения при большом числе обусловленности $\mu(A)$ и большом n с помощью последовательности (2), (3) и равенства (4) может натолкнуться на препятствие, состоящее в возможной потере вычислительной устойчивости при нахождении членов x_k последовательности (2). Однако для хорошо обусловленных систем и умеренных $\mu(A)$ метод (2)–(4) обладает достоинствами по сравнению с методами исключения, которые мы опишем в пп. 2, 3.

2. Произвольность формы задания оператора A. Очевидно, что в (2) используется только возможность по заданному элементу $y \in R^n$ находить элемент $z = Ay$ ($z \in R^n$), а также возможность по заданным произвольным $y \in R^n$, $z \in R^n$ вычислять их скалярное произведение, т. е. число (y, z). Поэтому не обязательно, чтобы система (1) была задана в каноническом виде

$$\sum a_{ij} x_j = f_i, \quad i = 1, 2, ..., n, \tag{5}$$

или была приведена к такому виду. К тому же нет необходимости хранить матрицу A, которая имела бы n^2 элементов, в то время как векторы $y \in R^n$, $z = Ay \in R^n$, записанные в координатной форме, задаются лишь n числами каждый.

3. Простота использования многопроцессорных вычислительных машин. Пусть система (1) задана в канонической форме (5), так что в записи (1) оператор A — матрица, а элементы $x, f \in R^n$ пространства R^n — наборы чисел $x = \begin{bmatrix} x_1 \\ ... \\ x_n \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ ... \\ y_n \end{bmatrix}$. Вычисление по

заданному $y = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$ каждой из n компонент вектора $Ay = z = \begin{bmatrix} z_1 \\ \dots \\ z_n \end{bmatrix}$ производится независимо. На многопроцессорной машине, допускающей одновременное осуществление многих арифметических операций, вычисление всех n компонент вектора Ay по заданному $y \in R^n$ может быть осуществлено одновременно и совершенно единообразно. Ускорение при использовании многопроцессорных ЭВМ может быть получено не только в случае задания системы (1) в канонической форме, но и в случае, если R^n состоит из сеточных функций, а значения сеточной функции $Ay \in R^n$ в каждой точке сетки вычисляются независимо (см., например, оператор $-\Delta^{(h)}\colon U^{(h)} \to U^{(h)}$, формулу (11) из § 1).

Заметим, что общее число ν арифметических операций при решении системы (5) методом сопряженных градиентов имеет порядок n^3, т. е. $\nu = O(n^3)$, как и в методе исключения Гаусса. Действительно, вычисление каждого члена x_{n+1} последовательности (2) требует применения матрицы A, т. е. $O(n^2)$ арифметических операций, а количество $k_0 \leqslant n$ таких членов может совпасть с n. Но вычисление Ay на многопроцессорной машине требует того же времени, что и вычисление одной компоненты при использовании однопроцессорной, так что время может быть сокращено в n раз.

Заметим, что при использовании метода Гаусса также возможно совмещение во времени многих арифметических операций.

§ 7. Конечные ряды Фурье и запись точного решения разностного аналога задачи Дирихле для уравнения Пуассона

Существуют методы, приспособленные для точного решения лишь узких классов СЛАУ, но зато более эффективные, чем универсальные. Среди них — представление решений СЛАУ в виде конечных рядов Фурье.

Пусть СЛАУ задана в виде

$$Ax = f, \quad x \in R^n, \quad f \in R^n, \quad A\colon R^n \to R^n, \tag{1}$$

причем $A = A^*$, $e_1, e_2, \dots, e_n$ — собственные векторы оператора A, образующие ортонормированный базис, а

$$\lambda_1, \lambda_2, \dots, \lambda_n, \quad \lambda_j \neq 0, \tag{2}$$

суть соответствующие собственные числа, $Ae_j = \lambda_j e_j$.

Каждый самосопряженный невырожденный оператор A имеет ортонормированный базис, состоящий из собственных векторов оператора A, а все его собственные значения вещественны и отличны от

нуля. Однако здесь мы предполагаем гораздо большее: считаем, что нам фактически известны этот ортонормированный базис, состоящий из собственных векторов, и соответствующие собственные значения. Это предположение весьма сужает класс допустимых самосопряженных операторов.

Итак, пусть выполнены предположения, сформулированные выше. Для вычисления решения системы (1) запишем заданное f и искомое x в виде конечных рядов Фурье:

$$f = F_1e_1 + F_2e_2 + \dots + F_ne_n, \quad F_j = (f, e_j), \tag{3}$$

$$x = X_1e_1 + X_2e_2 + \dots + X_ne_n. \tag{4}$$

Подставляя (3), (4) в уравнение (1), получаем

$$\sum_{j=1}^{n}(\lambda_j X_j)e_j = \sum_{j=1}^{n} F_j e_j. \tag{5}$$

Приравнивая коэффициенты при e_j в левой и правой частях (5), получаем значения коэффициентов X_j представления искомого решения в виде конечного ряда Фурье (4):

$$X_j = F_j/\lambda_j, \quad j = 1, 2, \dots, n. \tag{6}$$

Конкретизируем эту абстрактную схему применительно к вычислению решения $u^{(h)}$ разностного аналога задачи Дирихле для уравнения Пуассона в квадратной области (см. п. 3 § 1):

$$-\Delta^{(h)}u^{(h)} = f^{(h)}, \quad u^{(h)} \in U^{(h)}, \quad f^{(h)} \in F^{(h)}. \tag{7}$$

В этом случае удается указать собственные функции и собственные числа оператора $-\Delta^{(h)}$.

1. Ряды Фурье для сеточных функций. Рассмотрим множество всех вещественных функций $v = \{v_m\}$, определенных в точках $x_m = mh$ $(m = 0, 1, \dots, M)$, $Mh = 1$, обращающихся в нуль при $m = 0, M$. Совокупность этих функций с обычными операциями сложения и умножения их на вещественные числа образует линейное пространство. Размерность этого пространства есть $M - 1$, поскольку система функций

$$\widetilde{\psi}_m^{(k)} = \begin{cases} 0, & m \neq k, \\ 1, & m = k, \end{cases} \quad k = 1, 2, \dots, M - 1,$$

очевидным образом образует базис. Действительно, каждую функцию $v = (v_0, v_1, \dots, v_m)$, где $v_0 = v_M = 0$, можно единственным образом представить в виде линейной комбинации функций $\widetilde{\psi}^{(1)}, \widetilde{\psi}^{(2)}, \dots$ $\dots, \widetilde{\psi}^{(M-1)}$:

$$v = v_1\widetilde{\psi}^{(1)} + v_2\widetilde{\psi}^{(2)} + \dots + v_{M-1}\widetilde{\psi}^{(M-1)}.$$

Введем в рассматриваемом пространстве скалярное умножение, положив

$$(v, w) = h \sum_{m=0}^{M} v_m w_m. \tag{8}$$

Покажем, что система функций

$$\psi^{(k)} = \left\{ \sqrt{2} \sin \frac{k\pi m}{M} \right\}, \quad k = 1, 2, ..., M-1, \tag{9}$$

образует ортонормированный базис в рассматриваемом пространстве, т. е.

$$(\psi^{(k)}, \psi^{(r)}) = \begin{cases} 0, & k \neq r, \\ 1, & k = r, \end{cases} \qquad k, r = 1, 2, ..., M-1. \tag{10}$$

Для доказательства заметим, что

$$\sum_{m=0}^{M-1} \cos \frac{l\pi m}{M} = \frac{1}{2} \sum_{m=0}^{M-1} (e^{il\pi m/M} + e^{-il\pi m/M}) =$$

$$= \frac{1}{2}\,\frac{1 - e^{il\pi}}{1 - e^{i\pi/M}} + \frac{1}{2}\,\frac{1 - e^{-il\pi}}{1 - e^{i\pi/M}} = \begin{cases} 0, & l \text{ четно}, \\ 1, & l \text{ нечетно}, \end{cases} \qquad 0 < l < 2M.$$

Отсюда при $k \neq r$ получаем

$$(\psi^{(k)}, \psi^{(r)}) = 2h \sum_{m=0}^{M} \sin \frac{k\pi m}{M} \sin \frac{2\pi m}{M} = 2h \sum_{m=0}^{M-1} \sin \frac{k\pi m}{M} \sin \frac{r\pi m}{M} =$$

$$= h \sum_{m=0}^{M-1} \cos \frac{(k-r)\pi m}{M} - h \sum_{m=0}^{M-1} \cos \frac{2k\pi m}{M} = 0,$$

а при $k = r$

$$(\psi^{(k)}, \psi^{(r)}) = h \sum_{m=0}^{M-1} \cos 0 - h \sum_{m=0}^{M-1} \cos \frac{2k\pi m}{M} = hM - h \cdot 0 = 1.$$

Любая сеточная функция $v = (v_0, v_1, ..., v_M)$ разлагается по ортогональному базису (2) в сумму

$$v = c_1 \psi^{(1)} + c_2 \psi^{(2)} + ... + c_{M-1} \psi^{(M-1)},$$

или

$$v_m = \sqrt{2} \sum_{k=1}^{M-1} c_k \sin \frac{k\pi m}{M}, \tag{11}$$

где

$$c_k = (v, \psi^{(k)}) = \sqrt{2}\, h \sum_{m=0}^{M} v_m \sin\frac{k\pi m}{M}.$$

Ясно, что благодаря ортонормированности базиса (2) имеем

$$(v, v) = c_1^2 + c_2^2 + \ldots + c_{M-1}^2. \tag{12}$$

Сумма (11) и есть разложение сеточной функции $v = \{v_m\}$ в конечный ряд Фурье, а равенство (12) — точный аналог равенства Парсеваля в обычной теории рядов Фурье.

Совершенно аналогично можно рассмотреть конечные ряды Фурье для функций на сеточном квадрате. Рассмотрим сетку

$$x_m = mh, \quad y_n = nl, \quad 0 \leqslant mh \leqslant 1, \quad 0 \leqslant nh \leqslant 1,$$

причем $h = 1/M$, M — натуральное число. Совокупность вещественных функций $v = \{v_{mn}\}$, определенных в точках сетки и обращающихся в нуль в точках, лежащих на границе квадрата, образует линейное пространство. Введем в нем скалярное умножение

$$(v, w) = h^2 \sum_{n,m=0}^{M} v_{mn} w_{mn}.$$

В рассматриваемом линейном пространстве размерности $(M-1)^2$ система функций

$$\begin{gathered} \psi^{(k,l)} = 2 \sin\frac{k\pi m}{M} \sin\frac{l\pi n}{M}, \\ k = 1, 2, \ldots, M-1, \quad l = 1, 2, \ldots, M-1, \end{gathered} \tag{13}$$

образует ортонормированный базис

$$(\psi^{(k,l)}, \psi^{(r,s)}) = \begin{cases} 0, & \text{если } k \neq r \quad \text{или} \quad l \neq s, \\ 1, & \text{если } k = r \quad \text{и} \quad l = s. \end{cases}$$

Это следует из (3), если заметить, что

$$(\psi^{(k,l)}, \psi^{(r,s)}) = \left(2 \sum_{m=0}^{M} \sin\frac{k\pi m}{M} \sin\frac{r\pi m}{M}\right)\left(2 \sum_{n=0}^{M} \sin\frac{l\pi n}{M} \sin\frac{s\pi n}{M}\right) = \\ = (\psi^{(k)}, \psi^{(r)})(\psi^{(l)}, \psi^{(s)}).$$

Любая функция $v = \{v_{mn}\}$, обращающаяся в нуль на границе квадрата, разлагается в конечный двумерный ряд Фурье:

$$v_{mn} = 2 \sum_{k,\, l=1}^{M-1} c_{kl} \sin\frac{k\pi m}{M} \sin\frac{l\pi n}{M}, \tag{14}$$

где $c_{kl} = (v, \psi^{(k,l)})$. Справедливо равенство Парсеваля

$$(v, v) = \sum_{k=l=1}^{M-1} c_{kl}^2. \qquad (15)$$

2. Представление решения в виде конечного ряда Фурье. Можно проверить прямым вычислением справедливость равенств

$$\begin{gathered} -\Delta^{(h)}\psi^{(r,s)} = \frac{4}{h^2}\left(\sin^2\frac{r\pi}{2M} + \sin^2\frac{s\pi}{2M}\right)\psi^{(r,s)}, \\ r, s = 1, 2, ..., M-1, \end{gathered} \qquad (16)$$

где $-\Delta^{(h)}\colon U^{(h)} \to U^{(h)}$ определен в гл. 4, § 1.

Равенства (16) означают, что функции $\psi^{(r,s)}$, образующие ортонормированный базис в пространстве $U^{(h)}$, являются собственными функциями оператора $-\Delta^{(h)}\colon U^{(h)} \to U^{(h)}$, а числа

$$\lambda_{r,s} = \frac{4}{h^2}\left(\sin^2\frac{r\pi}{2M} + \sin^2\frac{s\pi}{2M}\right), \quad r, s = 1, 2, ..., M-1, \qquad (17)$$

суть отвечающие им собственные числа. В частности, отсюда следует, что оператор $-\Delta^{(h)}$ самосопряжен.

В соответствии с общей схемой решение задачи (7) запишется в виде

$$u_{m_1 m_2} = \sum_{r,s=1}^{M-1} \frac{F_{rs}}{\lambda_{rs}} \cdot 2\sin\frac{r\pi m_1}{M}\sin\frac{s\pi m_2}{M}, \qquad (18)$$

где

$$F_{rs} = 2h^2 \sum_{m_1,\, m_2=1}^{m_1,\, m_2=M} f_{m_1 m_2}\sin\frac{r\pi m_1}{M}\sin\frac{s\pi m_2}{M}. \qquad (19)$$

З а м е ч а н и е 1. В случае $N = 2^p$, где p — натуральное число, существует способ вычисления коэффициентов $F^{(r,s)}$ по формуле (19) и вычисления решения $u^{(h)}$ по формуле (18) за $O(M^2 \ln M)$ арифметических операций. Этот замечательный алгоритм называется *быстрым преобразованием Фурье* (см., например, [7, 12]).

Отметим, что

$$\lambda_{11} = \lambda_{11}(h) = \frac{8}{h^2}\sin^2\frac{\pi}{2M}.$$

Очевидно, что

$$\pi^2 \leqslant \lambda_{11} \leqslant 2\pi^2. \qquad (20)$$

Далее,

$$\lambda_{M-1,M-1} \sim 8/h^2. \qquad (21)$$

Поэтому число обусловленности оператора $-\Delta^{(h)}$

$$\mu(-\Delta^{(h)}) = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{\lambda_{M-1,M-1}}{\lambda_{11}} \qquad (22)$$

растет при $h \to 0$ как $O(h^{-2})$.

Заметим еще, что наибольшим собственным значением самосопряженного оператора $(-\Delta^{(h)})^{-1}$, которое совпадает с нормой этого оператора, является число λ_{11}^{-1}, которое превосходит в силу (20) величину π^2. Поэтому для решения задачи (7) при любой $f^{(h)}$ выполнено неравенство

$$\|u^{(h)}\| \leqslant \frac{1}{\pi^2}\,\|f^{(h)}\|. \tag{23}$$

Задача

Выписать с помощью конечного ряда Фурье решение разностного аналога задачи Дирихле для уравнения Пуассона

$$-\left(\frac{u_{m_1+1,m_2} - 2u_{m_1m_2} + u_{m_1-1,m_2}}{h^2} + \frac{u_{m_1,m_2+1} - 2u_{m_1m_2} + u_{m_1,m_2-1}}{h^2}\right) = f_{m_1m_2},$$

$$m_1 = 1, 2, ..., M-1, \quad m_2 = 1, 2, ..., M-1,$$

со следующими условиями на границе сеточного квадрата:

$$\begin{aligned} u_{0m_2} &= \varphi_{m_2}, & u_{Mm_2} &= \psi_{m_2}, & m_2 &= 1, 2, ..., M-1,\\ u_{m_10} &= \xi_{m_1}, & u_{m_1M} &= \eta_{m_1}, & m_1 &= 1, 2, ..., M-1, \end{aligned}$$

где $f_{m_1m_2}, \varphi_{m_2}, \psi_{m_2}, \xi_{m_1}, \eta_{m_1}$ — заданные функции своих аргументов.

Указание. Заметим, что решение поставленной задачи во внутренних точках совпадает с решением задачи, которая возникает при замене функций φ_{m_2}, ψ_{m_2}, ξ_{m_1}, η_{m_1} тождественно обращающимися в нуль и при одновременной замене значений $f_{m_1m_2}$ в приграничных точках сетки другими значениями $\widetilde{f}_{m_1m_2}$ (какими именно?). После этого решение внутри сеточной области записать в виде конечного ряда Фурье вида (18).

ГЛАВА 5

МЕТОДЫ ПОСЛЕДОВАТЕЛЬНЫХ ПРИБЛИЖЕНИЙ (ИТЕРАЦИОННЫЕ МЕТОДЫ) РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Задача отыскания точного решения уравнения

$$Ax = f, \quad f \in R^n, \quad x \in R^n, \tag{1}$$

где $A\colon R^n \to R^n$, A — некоторый линейный оператор, не диктуется, как правило, запросами приложений. В приложениях обычно допустимо использование приближенного решения, известного с достаточной для каждого данного приложения точностью. К тому же найти точное

решение, как правило, принципиально невозможно, так как входные данные (правая часть и сам оператор A) бывают известны не вполне точно, что приводит к неустранимой погрешности в результате. Кроме того, в силу конечной разрядности вычислительной машины неизбежны погрешности округления в процессе вычислений.

Поэтому во многих случаях для вычисления решения x уравнения (1) точным методам (методом Гаусса или другим) целесообразно предпочесть тот или иной метод последовательных приближений (итерационный метод). Каждый итерационный метод состоит в указании рекуррентного соотношения, которое по заданному произвольно нулевому приближению $x^{(0)}$ решения x позволяет вычислить первое, второе, вообще p-е $(p = 1, 2, ...)$ приближения $x^{(p)} \in R^n$ решения x.

Итерационный процесс должен быть построен так, чтобы последовательные приближения $x^{(p)}$ с ростом номера p стремились к решению x уравнения. Тогда для каждого $\varepsilon > 0$ можно указать номер $p = p(\varepsilon)$, чтобы выполнялось неравенство

$$||x - x^{(p)}|| < \varepsilon. \tag{2}$$

Задавая $\varepsilon > 0$ достаточно малым, можно воспользоваться p-м приближением $x^{(p)}$ в качестве приближенного решения с требуемой в данной задаче точностью.

Изложим некоторые итерационные методы и укажем условия, при которых эти методы целесообразно предпочесть точным или один другому.

§ 1. Методы простых итераций

Заметим, что СЛАУ

$$Ax = f \tag{1}$$

можно преобразовать к форме

$$x = (E - \tau A)x + \tau f, \tag{2}$$

причем новое уравнение (3) равносильно исходному при любом значении параметра $\tau (\tau > 0)$. Вообще $Ax = f$ многими способами можно заменить равносильной системой вида

$$x = Bx + \varphi, \quad x \in R^n, \quad \varphi \in R^n, \tag{3}$$

частным случаем которой является (2).

1. Общая схема метода простых итераций состоит в вычислении последовательности

$$x^{(p+1)} = Bx^{(p)} + \varphi, \quad p = 0, 1, ..., \tag{4}$$

при заданном произвольно значении $x^{(0)}$. Ниже мы укажем условия, при которых эта последовательность сходится к решению СЛАУ $Ax = f$.

Предварительно задержимся на частном случае итерационной формулы (4)

$$x^{(p+1)} = (E - \tau A)x^{(p)} + \tau f, \quad p = 0, 1, \ldots \tag{5}$$

Заметим, что для вычислений по этой формуле достаточно уметь по заданному $x^{(p)} \in R^n$ находить элемент $Ax^{(p)}$, получающийся в результате действия оператора A.

Таким образом, итерационный процесс вычисления решения СЛАУ $Ax = f$, в отличие от метода Гаусса, можно реализовать и в случае операторной формы задания СЛАУ, не выделяя какой-либо базис в R^n и не приводя систему к каноническому виду

$$\sum_{j=1}^{n} a_{ij}x_j = f_i, \quad i = 1, 2, \ldots, n. \tag{6}$$

В процессе вычислений по формуле (5) надо держать в памяти машины не n^2 чисел, составляющих матрицу оператора A в каком-либо базисе, а лишь n чисел, описывающих вектор $Ax^{(p)} \in R^n$. К тому же мы увидим, что для некоторых классов СЛАУ число арифметических действий для получения решения с разумной точностью итерационными методами много меньше, чем $O(n^3)$.

Теорема 1. *Пусть в R^n фиксирована некоторая норма, причем соответствующая норма оператора равносильной системы* (3) *оказалась меньше единицы*:

$$\|B\| = q < 1. \tag{7}$$

Тогда система (1) *имеет одно и только одно решение x; при любом $x^{(0)} \in R^n$ последовательность* (4) *сходится к решению x, причем погрешность p-го приближения* (*или p-й итерации*)

$$\varepsilon^{(p)} \equiv x - x^{(p)}$$

удовлетворяет оценке

$$\|\varepsilon^{(p)}\| = \|x - x^{(p)}\| \leqslant q^p\|x - x^{(0)}\| = q^p\|\varepsilon^{(0)}\|. \tag{8}$$

Тем самым норма погрешности $\|\varepsilon^{(p)}\|$ стремится к нулю с ростом p не медленнее геометрической прогрессии q^p.

Доказательство. В случае $\varphi = 0$ уравнение (3) не имеет решений, отличных от $x = 0$. В противном случае для решения $x \neq 0$, $\varphi = 0$ из (3) следовало бы, что

$$\|x\| = \|Bx\| \leqslant \|B\|\,\|x\| \leqslant q\|x\| < \|x\|,$$

т. е. $||x|| < ||x||$. В силу доказанного утверждения справедливо также утверждение о существовании и единственности решения СЛАУ (1) или (3) при произвольном φ.

Пусть x — решение системы (3). Зададим $x^{(0)} \in R^n$ произвольно. Вычитая из (3) равенство (4) почленно, получаем

$$\varepsilon^{(p+1)} = B\varepsilon^{(p)}, \quad p = 0, 1, \ldots$$

Отсюда имеем

$$||x - x^{(p)}|| = ||\varepsilon^{(p)}|| = ||B\varepsilon^{(p-1)}|| \leqslant q||\varepsilon^{(p-1)}|| \leqslant \\ \leqslant q^2||\varepsilon^{(p-2)}|| \leqslant \ldots \leqslant q^p||\varepsilon^{(0)}|| = q^p||x - x^{(0)}||. \square$$

Замечание. Условие (7) может нарушиться при каком-нибудь другом выборе нормы $||x||$. Однако сходимость сохранится, причем оценка (8) заменится оценкой

$$||\varepsilon^{(p)}||' \leqslant cq^p||\varepsilon^{(0)}||', \tag{8'}$$

где c — некоторая постоянная, зависящая от новой нормы, а значение q $(q < 1)$ прежнее.

Справедливость этого замечания следует из факта эквивалентности любых двух норм в конечномерном пространстве, состоящего в следующем. Если $||x||$, $||x||'$ — какие-нибудь нормы в R^n, то существуют такие числа $c_1 > 0$, $c_2 > 0$, что $c_1||x||' \leqslant ||x|| \leqslant c_2||x||'$, и c_1, c_2 от x не зависят. Легко видеть, что из (8) следует (8′) при $c = c_2/c_1$.

Пример 1. Пусть система (1) задана в каноническом виде (6), причем $A = \{a_{ij}\}$ с диагональным преобладанием, т. е. выполнены неравенства

$$|a_{ii}| > \sum_{j, j \neq i} |a_{ij}|, \quad i = 1, 2, \ldots, n.$$

В уравнении с номером i перенесем все слагаемые $a_{ij}x_j$ в правую часть и разделим это уравнение на a_{ii}. Возникнет система вида (3) с матрицей

$$B = \begin{bmatrix} 0 & b_{12} & b_{13} & \ldots & b_{1,n-1} & b_{1n} \\ b_{21} & 0 & b_{23} & \ldots & b_{2,n-1} & b_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ b_{n1} & b_{n2} & b_{n3} & \ldots & b_{n,n-1} & 0 \end{bmatrix}.$$

Очевидно, существует такое число q, что

$$\sum_{j=1}^{n} |b_{ij}| \leqslant q < 1,$$

а следовательно, первая норма $||B||_1 = \max\limits_i \sum\limits_j |b_{ij}|$ удовлетворяет оценке вида (7).

Зададим $x^{(0)}$ произвольно, но так, чтобы погрешность $\|\varepsilon^{(0)}\|_1 = \|x - x^{(0)}\|_1$ была, скажем, заведомо не больше единицы. И пусть нас устраивает ответ с погрешностью 10^{-3}. Тогда достаточно выбрать p так, чтобы выполнялось условие $q^p \leqslant 10^{-3}$. В случае, например, $q = 1/2$ достаточно взять $p = 10$ независимо от n.

Общее число арифметических операций будет $O(10n^2) = O(n^2)$, а не $O(n^3)$, как в методе Гаусса. Действительно, каждое применение матрицы B требует n^2 действий.

2. Необходимое и достаточное условие сходимости простых итераций.

Теорема 2. *Пусть B: $R^n \to R^n$, где R^n — n-мерное комплексное пространство. Последовательность простых итераций*

$$x^{(p+1)} = Bx^{(p)} + \varphi, \quad p = 0, 1, ...,$$

сходится (в произвольной норме) при произвольном $x^{(0)}$ в том и только том случае, если каждое собственное число λ_j оператора B строго меньше единицы по модулю:

$$|\lambda_j| \leqslant \rho < 1, \quad j = 1, 2, ..., n. \tag{9}$$

Доказательство. Сначала докажем достаточность условия (9) для сходимости. Пусть условие (9) выполнено. Заметим прежде всего, что тогда число $\lambda = 1$ не является собственным числом оператора B, и поэтому однородная система $By = 1 \cdot y$ имеет только тривиальное решение $y = 0$. Следовательно, система $x = Bx + \varphi$ всегда имеет решение x. Обозначим $x - x^{(p)} = \varepsilon^{(p)}$. Сходимость равносильна тому, что последовательность

$$\varepsilon^{(p+1)} = B\varepsilon^{(p)}, \quad p = 0, 1, ...,$$

сходится при произвольном ε_0 к нулю: $\|\varepsilon^{(p)}\| \to 0$. Фиксируем $\varepsilon^{(0)}$. Очевидно, что $\|\varepsilon^{(p)}\| \leqslant \|B\|^p \|\varepsilon^{(0)}\|$. Обозначим через $U(\lambda)$ сумму ряда векторных величин

$$U(\lambda) = \sum_{p=0}^{\infty} \frac{\varepsilon^{(p)}}{\lambda^p}.$$

Этот ряд заведомо сходится вне круга $|\lambda| > \|B\|$ на комплексной плоскости λ. Очевидно, что $\lambda U(\lambda) - \lambda\varepsilon_0 = BU(\lambda)$, или $U(\lambda) = -\lambda(B - \lambda E)^{(-1)}\varepsilon^{(0)}$. Из определения $U(\lambda)$ видно, что $\varepsilon^{(p)}$ является вычетом вектор-функции $\lambda^{p-1}U(\lambda)$:

$$\varepsilon^{(p)} = \frac{1}{2\pi i} \int\limits_{|\lambda|=r} \lambda^{p-1} U(\lambda)\, d\lambda = \frac{1}{2\pi r} \int\limits_{|\lambda|=r} \lambda^p (B - \lambda E)^{-1} \varepsilon^{(0)} d\lambda,$$

где r — любое число, которое больше $\|B\|$.

Пусть $|\lambda_j| \leqslant \rho < 1$. Тогда подынтегральная функция — аналитическая функция вне круга $|\lambda| > \rho$, так как оператор $(B - \lambda E)^{-1}$ существует при всех λ: $|\lambda| > \rho$. Поэтому можно деформировать контур интегрирования, выбирая $r = \rho + \varepsilon$, где $\varepsilon > 0$ произвольно, не изменяя интеграла. Отсюда

$$\|\varepsilon^{(p)}\| = \frac{1}{2\pi}\left| \int\limits_{|\lambda|=\rho+\varepsilon} \lambda^p (B - \lambda E)^{-1} \varepsilon^{(0)} d\lambda \right| \leqslant \\ \leqslant (\rho + \varepsilon)^p \max_{|\lambda|=\rho+\varepsilon} \|(B - \lambda E)^{-1}\| \, \|\varepsilon^{(0)}\|.$$

Выберем $\varepsilon > 0$ настолько малым, чтобы $\rho + \varepsilon$ было меньше единицы. Тогда правая часть последнего неравенства с ростом p стремится к нулю. Достаточность доказана.

Пусть условие (9) не выполнено, так что некоторое λ_k удовлетворяет неравенству $|\lambda_k| \geqslant 1$. Предположим, что вопреки утверждению теоремы сходимость $x^{(p)} \to x$ имеется при любом выборе $x^{(0)}$. Выберем x^0 так, чтобы $\varepsilon^{(0)} = e_k$, где e_k — собственный вектор оператора A, имеющий собственное число λ_k. Тогда $\varepsilon^{(p)} = B^p \varepsilon^{(0)} = B^p \varepsilon_k = \lambda_k^p e_k$. В силу $|\lambda_k| \geqslant 1$ эта последовательность не стремится к нулю. Получаем противоречие. □

Сделаем очевидное замечание, что если последовательность $x^{(p)}$ сходится, $x = \lim x^{(p)}$, то ее предел x является решением системы $x = Bx + \varphi$.

Отметим весьма интересное обстоятельство, которое встречается нам впервые. Задача вычисления $x = \lim x^{(p)}$ предельно хорошо обусловлена: результат вообще не зависит от входных данных, т. е. от выбора $x^{(0)}$. В то же время сходящийся в силу последней теоремы алгоритм вычисления последовательности $x^{(p)}$ может быть вычислительно очень неустойчивым. Неустойчивость может иметь место, если наряду с выполнением $\max|\lambda_j| = \rho < 1$ выполнено $\|B\| > 1$.

В случае $\|B\| \leqslant 1$ норма $\|\varepsilon^{(p)}\| = \|B^p \varepsilon^{(0)}\|$ монотонно убывает. В случае $\|B\| > 1$ существует $\varepsilon^{(0)}$ такое, что $\|\varepsilon^{(p)}\|$ сначала возрастает, а уж потом убывает. При этом высота "горба" может быть сколь угодно большой. Малая относительная погрешность округлений, допущенная при том p, где расположен максимум "горба", будет возрастать по норме: эта погрешность будет с ростом p также развиваться, проходя через максимум, и т. д. Вычислительная неустойчивость может оказаться уже при небольшом превышении $\|B\|$ над единицей и не очень больших n столь сильной, что расчет станет невозможным.

Строгое определение устойчивости метода простых итераций, классификацию видов возможной неустойчивости, примеры и теоремы см.: Рябенький В.С. Об устойчивости итерационных процессов// ДАН СССР. — 1970. — Т. 193, № 3, или [6, § 47].

3. Метод простых итераций в случае $A = A^* > 0$. Рассмотрим уравнение (3), т. е. $x = Bx + \varphi$, $x \in R^n$, в предположении, что R^n — евклидово пространство со скалярным произведением (x, y) и нормой $\|x\| = (x, x)^{1/2}$, а $B = B^*$ — самосопряженный оператор. Пусть ν_j $(j = 1, 2, ..., n)$ — собственные числа оператора B. Введем число

$$q = \max_j |\nu_j|.$$

Зададим произвольно начальное приближение $x^{(0)} \in R^n$ и построим последовательность простых итераций (4):

$$x^{(p+1)} = Bx^{(p)} + \varphi, \quad p = 0, 1, ... \tag{10}$$

Лемма 1. 1°. *Если $q < 1$, то уравнение $x = Bx + \varphi$ имеет решение x, а последовательные приближения $x^{(p)}$ в норме $\|x\| = (x, x)^{1/2}$ имеют погрешность $\|x - x^{(p)}\|$, которая с ростом p стремится к нулю и удовлетворяет оценке*

$$\|x - x^{(p)}\| \leqslant q^p \|x - x^{(0)}\|, \quad p = 0, 1, ..., \tag{11}$$

причем существует $x^{(0)}$ такое, что в (11) достигается точное равенство.

2°. *Пусть система $x = Bx + \varphi$ при данном φ имеет решение $x \in R^n$, но $q \geqslant 1$. Тогда существует начальное приближение $x^{(0)} \in R^n$ такое, что соответствующая последовательность простых итераций (10) не сходится к решению x.*

Доказательство. В случае $B = B^*$ евклидова норма оператора B совпадает с числом $q = \max |\nu_j|$. Поэтому утверждение 1° справедливо в силу теоремы 1.

Докажем 2°. Обозначим через $\varepsilon^{(p)} = x - x^{(p)}$ погрешность p-й итерации. Вычитая из тождества $\widetilde{x} = B\widetilde{x} + \varphi$ почленно равенство (10), получаем равенства $\varepsilon^{(p+1)} = B\varepsilon^{(p)}$.

Пусть $q \geqslant 1$, причем $q = |\nu_k|$. Зададим $x^{(0)}$ так, чтобы получить $\varepsilon^{(0)} = x - x^{(0)} = e_k$, где e_k — собственный вектор оператора B, имеющий собственное число λ_k. Очевидно, что в таком случае погрешность $\varepsilon^{(p)}$ не стремится к нулю с ростом p:

$$\varepsilon^{(p)} = B\varepsilon^{(p-1)} = ... = B^p \varepsilon^{(0)} = \lambda_k^p e_k. \square \tag{12}$$

Теорема 3. *Рассмотрим уравнение*

$$Ax = f, \quad A = A^* > 0, \tag{13}$$

относительно $x \in R^n$, где R^n — евклидово пространство. Пусть известны наименьшее $\lambda_{\min}$ и наибольшее $\lambda_{\max}$ собственные значения A. Зададим $\tau > 0$ и приведем уравнение (13) к виду

$$x = (E - \tau A)x + \tau f. \tag{14}$$

Зададим произвольное нулевое приближение $x^{(0)}$ *и рассмотрим последовательность простых итераций*

$$x^{(p+1)} = (E - \tau A)x^{(p)} + \tau f, \quad p = 0, 1, \ldots \tag{15}$$

1°. *Если* $\tau > 0$ *достаточно мало, а именно, удовлетворяет неравенствам*

$$0 < \tau < 2/\lambda_{\max}, \tag{16}$$

то последовательность $x^{(p)}$ *сходится к решению* x *уравнения* (13), *причем гарантировано убывание нормы погрешности* $\| x - x^{(p)} \|$ *при возрастании* p *в соответствии с оценкой*

$$\| x - x^{(p)} \| \leqslant q^p \| x - x^{(0)} \|, \quad p = 0, 1, \ldots \tag{17}$$

Здесь q *меньше единицы и выражается формулой*

$$q = q(\tau) = \max(|1 - \tau\lambda_{\min}|, |1 - \tau\lambda_{\max}|), \tag{18}$$

т. е. q — *наибольшее из двух чисел:* $|1 - \tau\lambda_{\min}|$, $|1 - \tau\lambda_{\max}|$, *каждое из которых при условии* (16) *строго меньше единицы.*

2°. *Пусть* τ — *произвольное число, удовлетворяющее* (16). *Существует начальное приближение* $x^{(0)}$, *при котором оценку* (17) *при выбранном* τ *улучшить нельзя, так как при этом* $x^{(0)}$ *соотношение* (17) *превращается в точное равенство.*

3°. *Если условие* (16) *нарушено так, что* $\tau \geqslant 2/\lambda_{\max}$ *или* $\tau < 0$, *то существует* $x^{(0)}$, *при котором последовательность* $x^{(p)}$ *не сходится с ростом* p *к решению* x.

4°. *Число* $q = q(\tau)$, *задаваемое формулой* (18), *принимает наименьшее значение* $q_{\text{опт}} = q(\tau_{\text{опт}})$, *если* $\tau = \tau_{\text{опт}} = 2/(\lambda_{\min} + \lambda_{\max})$. *В этом случае число* q *принимает значение*

$$q = q_{\text{опт}} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\mu(A) - 1}{\mu(A) + 1}, \tag{19}$$

где $\mu(A) = \lambda_{\max}/\lambda_{\min}$ — *число обусловленности оператора* A.

Доказательство. Оно опирается на лемму 1. В этой лемме полагаем $B = E - \tau A$. Заметим, что если $A = A^*$, то оператор $B = E - \tau A$ также самосопряженный:

$$(Bx, y) = ((E - \tau A)x, y) = (x, y) - \tau(Ax, y) = \\ = (x, y) - \tau(x, Ay) = (x, (E - \tau A)y) = (x, By).$$

Пусть $\lambda_1, \lambda_2, \ldots, \lambda_n$ — собственные числа оператора A, расположенные в порядке неубывания:

$$0 < \lambda_{\min} = \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n = \lambda_{\max}, \tag{20}$$

$e_1, e_2, \ldots, e_n$ — соответствующие собственные векторы, образующие ортонормированный базис, $Ae_j = \lambda_j e_j$. Тогда те же векторы являются

собственными для оператора B, причем соответствующие собственные числа есть

$$\nu_j = \nu_j(\tau) = 1 - \tau\lambda_j, \quad j = 1, 2, ..., n,$$
$$Be_j = (E - \tau A)e_j = e_j - \tau\lambda_j e_j = (1 - \tau\lambda_j)e_j = \nu_j e_j, \tag{21}$$
$$j = 1, 2, ..., n.$$

Очевидно, что в силу (20) собственные числа ν_j расположены в порядке невозрастания (рис. 11):

$$\nu_1 \geqslant \nu_2 \geqslant ... \geqslant \nu_n. \tag{22}$$

Из рис. 11 видно, что наибольшим среди чисел $|\nu_j|$ $(j = 1, 2, ..., n)$ является либо $|\nu_1| = |1 - \tau\lambda_{\min}|$, либо $|\nu_n| = |1 - \tau\lambda_{\max}|$, так что условие

$$q = \max_j |\nu_j| < 1 \tag{23}$$

леммы 1 совпадает с условием

$$q = \max\{|1 - \tau\lambda_{\min}|, |1 - \tau\lambda_{\max}|\} < 1. \tag{24}$$

Условие $q < 1$ при $\tau > 0$ выполняется в том и только том случае, если точка ν_n на рис. 11 лежит правее точки -1 (т. е. если $1 - \tau\lambda_{\max} > -1$). Это означает, что наряду с условием $\tau > 0$ выполняется и второе неравенство (16). Если $\tau \geqslant 2/\lambda_{\max}$, то $q > 1$. Если $\tau < 0$, то $\nu_j = 1 - \tau\lambda_j = 1 + |\tau|\lambda_j > 1$, так что $q = \max|\nu_j| > 1$. Итак, (16) равносильно требованию (9) леммы 1 в случае $B = E - \tau A$. Таким образом, доказаны утверждения 1°–3°.

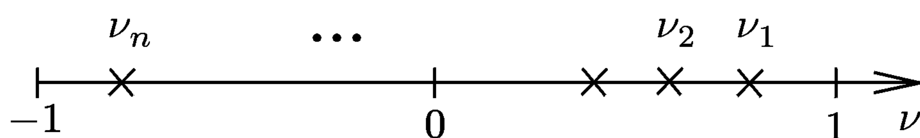


Рис. 11

Докажем теперь утверждение 4°. При очень маленьких τ $(\tau > 0)$ все точки $\nu_1, \nu_2, ..., \nu_n$ расположены на рис. 11 левее точки 1, но вблизи нее, так что $\max|\nu_j| = \nu_1 = 1 - \tau\lambda_{\min}$. При увеличении τ все точки смещаются влево: число $q = \max|\nu_j| = 1 - \tau\lambda_1$ уменьшается. Так будет происходить до тех пор, пока не наступит равенство $|\nu_n| = \nu_1$. При дальнейшем увеличении τ число ν_1 будет продолжать уменьшаться. Однако теперь окажется, что $|\nu_n| > \nu_1$, так что станет $q = \max|\nu_j| = |\nu_n|$, но число $|\nu_n|$ с ростом τ будет расти. Наименьшее значение $q = q_{\text{опт}}$ будет при том значении $\tau = \tau_{\text{опт}}$, при котором $|\nu_n| = \nu_1$, или $-(1 - \tau\lambda_{\text{n}}) = 1 - \tau\lambda_1$, т. е. при $\tau_{\text{опт}} = 2/(\lambda_{\min} + \lambda_{\max})$. Очевидно, что при таком значении $\tau = \tau_{\text{опт}}$ число $q = q_{\text{опт}} = q(\tau_{\text{опт}})$ есть

$$q = \frac{\mu(A) - 1}{\mu(A) + 1} = \nu_1 = 1 - \tau_{\text{опт}}\lambda_{\min} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}. \square$$

Очевидно, что чем ближе $\mu(A)$ к единице, тем ближе к нулю $q_{\text{опт}}$ и тем быстрее в силу (17) убывает погрешность. С ростом числа обусловленности $\mu(A)$ число $q_{\text{опт}}$ увеличивается (оставаясь меньше единицы) и сходимость замедляется.

При оптимальном выборе $\tau = \tau_{\text{опт}}$ справедлива оценка

$$\|\varepsilon^{(p)}\| \leqslant \left(\frac{1-\xi}{1+\xi}\right)^p \|\varepsilon^{(0)}\|, \quad \xi = \frac{a}{b} = \frac{\lambda_{\min}}{\lambda_{\max}}.$$

В силу леммы 1 существуют такие A, $\varepsilon^{(0)}$, что эта оценка достигается. Поэтому для того чтобы гарантировать оценку

$$\|\varepsilon^{(p)}\| \leqslant \varepsilon \|\varepsilon^0\|, \quad p = 0, 1, \ldots, \tag{25}$$

при заданном $\varepsilon > 0$, число p необходимо и достаточно выбрать из условия

$$\left(\frac{1-\xi}{1+\xi}\right)^p \leqslant \varepsilon, \quad p \geqslant -\frac{\ln \varepsilon}{\ln(1+\xi) - \ln(1-\xi)}.$$

Укажем более обозримую оценку для p. Заметим, что

$$\ln(1+\xi) - \ln(1-\xi) = 2\xi \sum_{k=0}^{\infty} \frac{\xi^{2k}}{2k+1},$$

$$1 \leqslant \sum_k \frac{\xi^{2k}}{2k+1} \leqslant \frac{1}{1-\xi^2}.$$

Поэтому для гарантированной оценки (25) достаточно, чтобы значение p удовлетворяло оценке

$$p \geqslant -\frac{1}{2} \ln \varepsilon \cdot \mu, \quad \mu = \frac{1}{\xi} = \frac{b}{a}, \tag{26}$$

и необходимо, чтобы

$$p \geqslant -\frac{1}{2} \ln \varepsilon \cdot (1-\xi^2)\mu. \tag{27}$$

Замечание. Во многих случаях (например, при приближенной замене некоторых эллиптических краевых задач разностными) оператор $A\colon R^n \to R^n$ возникающей линейной системы оказывается положительно определенным и самосопряженным ($A = A^* > 0$) в смысле некоторого естественного скалярного умножения. Однако обычно не удается точно указать его наибольшее и наименьшее собственные значения. Удается указать лишь оценки границ спектра, т. е. такие числа a, b, чтобы выполнялись неравенства

$$0 < a \leqslant \lambda_{\min} \leqslant \lambda_{\max} \leqslant b.$$

В этом случае также можно воспользоваться методом простых итераций (15).

Зная вместо $\lambda_{\min}$, $\lambda_{\max}$ лишь границы a, b спектра, можно воспользоваться значением $\tau' = 2/(a+b)$. При этом вместо

$$q_{\text{опт}} = (\lambda_{\max} - \lambda_{\min})/(\lambda_{\max} + \lambda_{\min})$$

в гарантированной оценке (17) будет фигурировать число

$$q' = \max\{|1 - \tau'\lambda_{\min}|, |1 - \tau'\lambda_{\max}|\},$$

которое больше, чем $q_{\text{опт}}$. Как показано выше, при любом выборе τ оценка (17) при некотором $x^{(0)}$ превращается в точное равенство, причем $q = q(\tau)$ определяется формулой (18). Поэтому при $\tau = \tau' \neq \tau_{\text{опт}}$ получим неулучшаемую оценку (17), в которой будет $q = q(\tau') > q_{\text{опт}}$. Сходимость окажется тем более медленной, чем грубее известны границы a, b спектра.

Пример 2. Применим метод простых итераций к вычислению решения разностного аналога задачи Дирихле для уравнения Пуассона $-\Delta^{(h)}u^{(h)} = f^{(h)}$, сформулированной в § 7 из гл. 4.

Формула (15) примет вид

$$u^{(p+1)} = (E + \tau\Delta^{(h)})u^{(p)} + \tau f, \quad p = 0, 1, \ldots$$

Собственные числа оператора $-\Delta^{(h)}$ найдены в (17) из § 7 гл. 4. Там же введено естественное скалярное умножение в пространстве $U^{(h)}$, где действует оператор $-\Delta^{(h)}$, установлена его самосопряженность и найдено число обусловленности $\mu(-\Delta^{(h)}) = O(h^{-2})$.

Поэтому в силу (25) при $\tau_{\text{опт}}$ (укажите это $\tau_{\text{опт}}$) количество итераций для уменьшения погрешности в e раз составит $p \approx \frac{1}{2}\mu(-\Delta^{(h)}) = O(h^{-2})$. Каждая итерация требует $O(h^{-2})$ арифметических операций; их общее число $O(h^{-4})$.

В § 7 гл. 4 мы выписали точное решение задачи (20) в виде конечного ряда Фурье. Однако в случае непрямоугольной области или в случае, если вместо разностного аналога уравнения Пуассона рассматривался бы разностный аналог уравнения

$$\frac{\partial}{\partial x}\left(a\,\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(b\,\frac{\partial u}{\partial y}\right) = f \tag{28}$$

с переменными коэффициентами $a = a(x, y) > 0$, $b(x, y) > 0$, мы не знали бы собственных функций и собственных чисел задачи и не могли бы воспользоваться конечными рядами Фурье. В то же время алгоритм простых итераций можно было бы построить и в этом случае вполне аналогично тому, как это сделано выше. Нужно лишь, чтобы разностное уравнение было самосопряженным и чтобы были известны не слишком грубые границы a, b спектра оператора задачи.

В дальнейшем мы укажем для уравнений $Ax = f$ $(A = A^* > 0)$ с плохо обусловленным оператором A гораздо более эффективные итерационные алгоритмы, чем метод простой итерации.

4. Переход от $Ax = f$ ($A = A^* > 0$) к системе, которая лучше обусловлена, с помощью энергетически эквивалентного оператора. Мы видели, что скорость сходимости метода простых итераций (как и других, которые будут указаны в § 2) тем выше, чем число обусловленности системы $\mu(A)$ меньше, т. е. чем это число ближе к единице.

В случае плохо обусловленной системы $Ax = f$ иногда удается перейти к равносильной системе с оператором, который имеет меньшее число обусловленности, а затем решать эту систему методом итераций. Изложим этот прием.

Пусть $B = B^* > 0$ — пока произвольный оператор. Умножим обе части уравнения $Ax = f$ на B^{-1}. Получим равносильное уравнение

$$Cx = g, \quad C = B^{-1}A, \quad g = B^{-1}f. \tag{29}$$

Оператор C уже не является, вообще говоря, самосопряженным.

Введем новое скалярное умножение $[x, y]_B = (Bx, y)$. Оператор C оказывается в смысле нового скалярного умножения самосопряженным, т. е. $[(x, y)]_B = [x, Cy]_B$, а также положительно определенным, т. е. $[Cx, x]_B > 0$, если $x \neq 0$. Проверим это:

$$[Cx, y]_B = (BCx, y) = (BB^{-1}Ax, y) = (Ax, y) = (x, Ay) =$$
$$= (B^{-1}Bx, Ay) = (Bx, B^{-1}Ay) = [x, Cy]_B,$$
$$[Cx, x]_B = (BCx, x) = (Ax, x) > 0, \quad x \neq 0.$$

В нашем переходе от уравнения $Ax = f$ к уравнению (29) выбор оператора пока произволен. В частности, если положить $B = A$, то оператор $C = B^{-1}A$ окажется единичным, и решение x будет получено. Однако применение оператора A^{-1} равносильно точному решению уравнения $Ax = f$, которого мы как раз и хотим избежать за счет итерационного процесса, требующего умения вычислять при заданном z вектор Az, но не $A^{-1}f$. Поэтому имеет смысл выбирать оператор B лишь среди тех, для которых вычисление $B^{-1}z$ по заданному z существенно проще, чем вычисление $A^{-1}z$. Если при этом удается выбрать B так, чтобы он был “похож” на оператор A, то можно надеяться, что оператор $B^{-1}A$ будет “похож” на единичный, а его собственные числа $\lambda_{\min}$, $\lambda_{\max}$ и число обусловленности $\mu(C)$ будут “ближе” к единице.

Теорема 4. *Пусть $B = B^* > 0$ и при заданных числах $\gamma_1 > 0$, $\gamma_2 > 0$ и всех $x \in R^n$ справедливы неравенства*

$$\gamma_1(Bx, x) \leqslant (Ax, x) \leqslant \gamma_2(Bx, x). \tag{30}$$

Тогда собственные числа $\lambda_{\min}(C)$, $\lambda_{\min}(C)$ и число обусловленности $\mu_B(C)$ оператора $C = B^{-1}A$ удовлетворяют неравенствам

$$\gamma_1 \leqslant \lambda_{\min}(C) \leqslant \lambda_{\max}(C) \leqslant \gamma_2, \quad \mu_B(C) \leqslant \gamma_2/\gamma_1. \tag{31}$$

Доказательство. Из курса линейной алгебры известно (и легко видеть непосредственно), что

$$\lambda_{\min}(C) = \min_x \frac{[Cx,x]_B}{[x,x]_B} = \min_x \frac{(Ax,x)}{(Bx,x)},$$
$$\lambda_{\max}(C) = \max_x \frac{[Cx,x]_B}{[x,x]_B} = \max_x \frac{(Ax,x)}{(Bx,x)}.$$

Отсюда в силу (30) следуют неравенства (31). □

Операторы A, B, удовлетворяющие неравенствам (30), принято называть *эквивалентными по спектру* или *энергетически эквивалентными* с константами эквивалентности γ_1, γ_2.

Переход от $Ax = f$ к $Cx = g$ имеет смысл и существенно улучшает число обусловленности, если

$$\mu_B(C) \leqslant \gamma_2/\gamma_1 \ll \mu(A).$$

При таком переходе увеличивается по сравнению с (17) скорость убывания погрешности $\varepsilon^{(p)} = x - x^{(p)}$ в норме $\|\cdot\|_B$:

$$\|\varepsilon^{(p)}\|_B = \|x - x^{(p)}\|_B \leqslant q_B^p \|x - x^{(0)}\|_B,$$
$$q_B = \frac{\mu_B(C) - 1}{\mu_B(C) + 1}.$$

Тогда при том же значении p будет $q_B^p \ll q^p$, так как

$$q_B \ll q, \quad q = \frac{\mu(A) - 1}{\mu(A) + 1}.$$

Типичная ситуация, в которой использование эквивалентных по спектру операторов дает большой эффект, впервые выделена и изучена Е.Г. Дьяковым в начале 60-х годов и состоит в следующем. Пусть система алгебраических уравнений

$$A_n x = f, \quad f \in \mathsf{R}^n, \quad x \in \mathsf{R}^n,$$

где $A_n\colon \mathsf{R}^n \to \mathsf{R}^n$, возникла при дискретизации краевой задачи для эллиптического дифференциального уравнения. Размерность n пространства R^n тем выше, чем более точное приближение A_n исходной задачи мы используем при дискретизации. Таким образом, мы имеем дело с последовательностью пространств R^n $(n \to \infty)$, которые будем считать эвклидовыми со скалярным умножением $(x,y)^{(n)}$.

Пусть $A_n\colon \mathsf{R}^n \to \mathsf{R}^n$ — последовательность операторов, причем $A_n = A_n^* > 0$, а $A_n x = f$ $(x, f \in \mathsf{R}^n)$ — последовательность уравнений. Пусть числа обусловленности $\mu(A_n)$ растут с ростом n так, что $\mu(A_n) \sim n^s$ $(s > 0, s = \mathrm{const})$. Тогда при отыскании решения с точностью $\varepsilon > 0$ итерациями в соответствии с (27) потребуется $O(n^s \ln \varepsilon)$ итераций.

Пусть $B_n\colon R^n \to R^n (B_n = B_n^* > 0)$ энергетически эквивалентно с оператором A_n с константами эквивалентности $\gamma_1 > 0$, $\gamma_2 > 0$, которые не зависят от n. Тогда $\mu_{B_n}(C_n) \leqslant \gamma_2/\gamma_1 = \mathrm{const}$ $(C_n = B_n^{-1}A_n)$.

Перейдем к последовательности уравнений

$$C_n x = g_n, \quad C_n = B_n^{-1}A_n, \quad g_n = B_n^{-1}f^{(n)}.$$

Количество итераций для уменьшения погрешности в ε раз:

$$\|x - x^{(p)}\|_{B_n} \leqslant \varepsilon \|x - x^{(0)}\|_{B_n}, \tag{32}$$

в силу ограниченности $\mu_{B_n}(C_n) \leqslant \gamma_2/\gamma_1$ не будет возрастать с ростом n и будет $O(\ln\varepsilon)$.

Пусть нормы

$$\|x\| = [(x,x)^{(n)}]^{1/2}, \quad \|x\|_{B_n} = \{[x,x]_{B_n}\}^{1/2}$$

связаны неравенствами

$$n^{-l}\|x\|_{B_n} \leqslant \|x\| \leqslant n^l\|x\|_{B_n}, \quad l \geqslant 0, \quad l = \mathrm{const}.$$

Тогда для достижения оценки

$$\|x - x^{(p)}\| \leqslant \varepsilon\,\|x - x^{(0)}\|$$

в итерационном процессе для уравнения $C_n x = g^{(n)}$ достаточно, чтобы выполнялось неравенство

$$\|x - x^{(p)}\|_{B_n} \leqslant \frac{\varepsilon}{n^l}\,\|x - x^{(0)}\|_{B_n}. \tag{33}$$

Это неравенство получено из (32) заменой ε на εn^{-l}, так что потребуется лишь $O\left(\ln\dfrac{\varepsilon}{n^l}\right) = O(\ln n - \ln\varepsilon)$ итераций, в то время как без использования перехода от $A_n x = f$ к $C_n x = g^{(n)}$ требовалось, напомним, $O(n^l \ln\varepsilon)$ итераций.

5. Масштабирование как средство улучшения числа обусловленности. Пусть требуется решить систему

$$Cx = \varphi, \tag{34}$$

заданную в каноническом виде: $C = \{c_{ij}\}$, $\det C \neq 0$. Можно перейти к системе

$$Ax = f, \quad A = C * C, \quad f = C * \varphi,$$

причем $A = A^* > 0$, а затем воспользоваться методом итераций. Скорость их сходимости зависит от числа обусловленности $\mu(A)$. Легко показать, что в евклидовой норме, соответствующей скалярному умножению в пространстве R^n, $\mu(A) = \mu^2(C)$. Прием масштабирования для уменьшения числа обусловленности заключается в следующем. Каждое из скалярных уравнений, составляющих (34), умножается на

такой множитель, чтобы наибольший коэффициент нового уравнения оказался равным единице. От этой новой, масштабированной системы переходим к равносильной системе с симметричной и положительной определенной матрицей.

Прием масштабирования уменьшает число обусловленности исходной системы. В силу равенства $\mu(A) = \mu^2(C)$ уменьшается также число $\mu(A)$, определяющее скорость сходимости итераций.

Задачи

1. Пусть известно, что собственные значения оператора $A\colon \mathbb{R}^{100} \to \mathbb{R}^{100}$ суть

$$\lambda_k = k^2, \quad k = 1, 2, ..., 100.$$

Для приближенного вычисления решения системы $Ax = f$ произвольно задается начальное приближение $x^{(0)} \in \mathbb{R}^{100}$, а последующие приближения вычисляются по формулам

$$x^{(p+1)} = (E - \tau_p A)x^{(p)} + \tau_p f, \quad p = 0, 1, ..., \tag{35}$$

где τ_p — некоторые положительные числа.

Укажите такой набор итерационных параметров $\tau_0, \tau_1, ..., \tau_{99}$, чтобы приближение $x^{(100)}$ совпадало с точным решением x системы $Ax = f$.

2. Пусть в предыдущей задаче итерационные параметры τ_p выбраны по формуле

$$\tau_p = \frac{1}{(p+1)^2}, \quad p = 0, 1, ..., 99.$$

а) Проверить, что тогда $x^{(100)} = x$.

б) При реализации алгоритма (35) на компьютере с допустимым десятичным порядком чисел, скажем, до 10, вычисления натолкнутся на препятствие, которое состоит в том, что появятся числа слишком большого порядка. Объяснить механизм возникновения этого явления.

3. Пусть в задаче 1 итерационные параметры τ выбраны по формуле

$$\tau_p = \frac{1}{(101-p)^2}, \quad p = 1, 2, ..., 100.$$

а) Проверить, что тогда $x^{(100)} = x$.

б) При реализации этого алгоритма на компьютере с заданным количеством разрядов, скажем, 10, возникает большая погрешность. Объяснить механизм, препятствующий расчету.

§ 2. Метод Чебышева и метод сопряженных градиентов

Для системы вида

$$Ax = f, \quad A = A^* > 0, \quad A\colon \mathbb{R}^n \to \mathbb{R}^n, \tag{1}$$

укажем два итерационных алгоритма для вычисления решения $x = \overline{x}$, в которых заданная точность достигается при меньшем объеме вычислений, чем в методе простых итераций, описанном в § 1. Обсудим также условия, при которых один из этих двух методов предпочти-

тельнее другого. Обоснования указанных ниже алгоритмов см., например, в [20].

В силу $A = A^* > 0$ собственные числа λ_j $(j = 1, 2 ..., n)$ оператора строго положительны. Будем считать, что они занумерованы в порядке возрастания и что известны такие числа $a > 0$, $b > 0$, что выполнено

$$a \leqslant \lambda_1 \leqslant ... \leqslant \lambda_n \leqslant b. \tag{2}$$

Числа a, b называются *границами спектра* оператора A. Если $a = \lambda_1$, $b = \lambda_n$, то a, b называют *точными границами спектра.*

1. Метод Чебышева. Зададим произвольно нулевое приближение x^0 и будем вычислять последующие приближения по формулам

$$x^{(1)} = \big(E - \tau A\big)x^{(0)} + \tau f, \tag{3}$$

$$x^{(p+1)} = \alpha_{p+1}(E - \tau A)x^{(p)} + (1 - \alpha_{p+1})x^{(p-1)} + \tau\alpha_{p+1}f,$$
$$p = 1, 2, ...,$$

где τ и α_k заданы формулами

$$\tau = \frac{2}{a+b}, \quad \alpha_1 = 2, \quad \alpha_{p+1} = \frac{4}{4 - \rho^2\alpha_p}, \quad p = 1, 2, ..., \quad \rho = \frac{b+a}{b-a}. \tag{4}$$

Можно показать, что погрешность $\varepsilon^{(p)} = \overline{x} - x^{(p)}$, $||\varepsilon^{(p)}||^2 = (\varepsilon^{(0)}, \varepsilon^{(p)})$, приближения $x^{(p)}$ удовлетворяет оценке

$$||\varepsilon^{(p)}|| \leqslant \frac{2p^q}{1+q^{2p}}\,||\varepsilon^{(0)}||, \quad q = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{a}{b}, \quad p = 0, 1, ... \tag{5}$$

Для уменьшения первоначальной погрешности в ε^{-1} раз, т. е. для достижения оценки $||\varepsilon^{(p)}|| \leqslant \varepsilon||\varepsilon^{(0)}||$, достаточно, чтобы p было выбрано из условия

$$p \geqslant -\frac{1}{2}\Big(\ln\frac{\varepsilon}{2}\Big)\sqrt{\frac{b}{a}}.$$

Это число шагов p примерно в $\nu = \frac{1}{2}\sqrt{\frac{b}{a}}$ раз меньше, чем число шагов, которого требует для достижения той же оценки $||\varepsilon^{(p)}|| \leqslant \varepsilon||\varepsilon^{(0)}||$ метод простой итерации. Экономия тем больше, чем больше число $b/a = \xi^{-1}$. В случае, если a, b — точные границы спектра $\frac{b}{a} = \frac{\lambda_n}{\lambda_1} = \mu(A)$ и $\nu = \frac{1}{2}\sqrt{\mu(A)}$, экономия тем больше, чем больше число обусловленности системы (1). Отметим, что итерационный алгоритм (3) вычислительно устойчив.

Конструкция метода (3), доказательство оценки (5), доказательства свойства вычислительной устойчивости алгоритма (3) и других его свойств основаны на использовании многочленов Чебышева и некоторых других связанных с ними многочленов. Поэтому алгоритм (3) называют *трехслойным чебышевским алгоритмом.*

Существует и используется на практике также так называемый

двухслойный чебышевский итерационный алгоритм для уравнения (1). Для вычисления приближения $x^{(p+1)}$ этот алгоритм использует и требует хранения в памяти машины только $x^{(p)}$, т. е., в отличие от (3), $x^{(p-1)}$ хранить не нужно. Однако этот алгоритм требует дополнительной заботы об устойчивости счета, и мы его не излагаем.

2. Метод сопряженных градиентов. Формулы для вычисления приближений $x^{(p+1)}$ методом сопряженных градиентов см. в § 6 гл. 4. В этих формулах не используются какие-либо границы a, b спектра, что является существенным преимуществом перед методом Чебышева.

Гарантированная оценка скорости убывания погрешности зависит от числа обусловленности $\mu(A)$ и не медленнее, чем в методе Чебышева (3), построенного в случае известных a, b, т. е. в случае известных точных границ спектра.

В отличие от метода Чебышева при использовании метода сопряженных градиентов может возникнуть вычислительная неустойчивость, которая проявляется тем сильнее, чем больше число обусловленности $\mu(A)$. Наиболее благоприятна для применения метода сопряженных градиентов ситуация, когда известно, что число обусловленности $\mu(A)$ невелико, но границы спектра неизвестны, а порядок n системы много больше того числа итераций p, при котором погрешность $\varepsilon^{(p)}$ удовлетворяет поставленному перед вычислителем требованию точности.

Скорость сходимости метода Чебышева и метода сопряженных градиентов можно увеличить, если преобразовать систему (1) к системе вида

$$Cx = \varphi, \quad C = B^{-1}A,$$

$B = B^* > 0$, уменьшив при этом число обусловленности. В этом случае полностью сохраняются построения и рассуждения из п. 4 § 1.

ГЛАВА 6

ПЕРЕОПРЕДЕЛЕННЫЕ СЛАУ. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

§ 1. Примеры задач, приводящих к переопределенным СЛАУ

1. Задача обработки экспериментальных данных — эмпирические формулы. Пусть известно, что величина y является некоторой функцией от аргумента t, причем в результате измерений

получена таблица значений $y_k = y(t_k)$ $(k = 1, 2, 3, 4)$ этой функции (рис. 12).

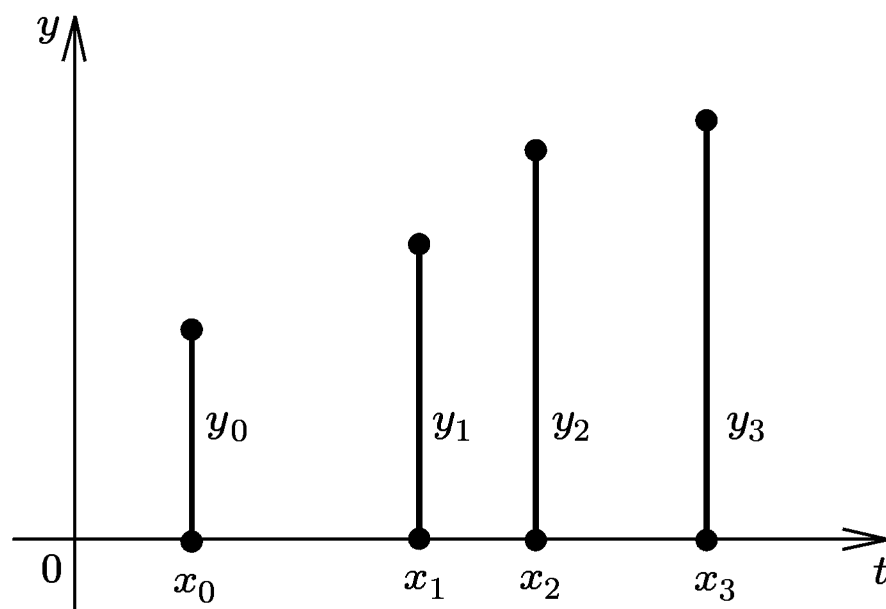


Рис. 12

Полученные измерения позволяют приближенно считать, что зависимость $y = y(t)$ является линейной, т. е. имеет вид

$$y = x_1 t + x_2, \tag{1}$$

где x_1, x_2 — некоторые числа. Числа x_1, x_2 в эмпирической формуле (1) хотелось бы подобрать так, чтобы при $t = t_k$ $(k = 1, 2, 3, 4)$ формула (1) давала те значения, которые получены измерениями, т. е. чтобы выполнялись уравнения

$$\begin{aligned} x_1 t_1 + x_2 &= y_1, \\ x_1 t_2 + x_2 &= y_2, \\ x_1 t_3 + x_2 &= y_3, \\ x_1 t_4 + x_2 &= y_4. \end{aligned} \tag{2}$$

Получилась система четырех линейных уравнений относительно двух неизвестных x_1, x_2. Эта переопределенная система не имеет (классического) решения, так как не существует прямой, проходящей через все четыре экспериментальные точки (см. рис. 12).

За значения x_1, x_2, входящие в формулу (1), можно принять ту пару чисел, для которой невязки

$$[x_1 t_k + x_2 - y_k], \quad k = 1, 2, 3, 4,$$

будут как можно меньше. Можно, например, ввести функцию

$$\Phi(x_1, x_2) = \sum_{k=1}^{4} |x_1 t_k + x_2 - y_k|^2, \tag{3}$$

равную сумме квадратов невязок, и принять за обобщенное решение x_1, x_2 системы (2) ту пару чисел x_1, x_2, для которой $\Phi(x_1, x_2)$

принимает наименьшее значение. Для вычисления этих чисел x_1, x_2 получим систему

$$\partial\Phi/\partial x_1 = 0, \quad \partial\Phi/\partial x_2 = 0 \tag{4}$$

двух уравнений, имеющую (обычное, классическое) решение x_1, x_2.

Выбор функции $\Phi(x_1, x_2)$, от которой зависит, какую именно пару чисел x_1, x_2 принять за обобщенное решение системы (2), т. е. за коэффициенты эмпирической формулы (1), содержит произвол. Если бы, например, мы сочли разумным придать каждому измерению свой вес b_k $(k = 1, 2, 3, 4)$, то вместо (3) было бы естественно принять функцию

$$\Phi(x_1, x_2) = \sum_{k=1}^{4} b_k (x_1 t_k + x_2 - y_k)^2. \tag{5}$$

Соответственно изменилась бы и та пара чисел x_1, x_2 (обобщенное решение системы (2)), которая придает наименьшее значение функции $\Phi(x_1, x_2)$. Система (4) и ее решение x_1, x_2 зависят от b_1, b_2, b_3, b_4.

Определение обобщенного решения системы (2) с помощью минимизации функции $\Phi(x_1, x_2)$ от квадратов невязок есть пример метода наименьших квадратов для определения обобщенного решения переопределенной системы (2).

Можно было бы ввести меру невязки $\Phi(x_1, x_2)$, используя вместо (5) формулу

$$\Phi(x_1, x_2) = \sum_{k=1}^{4} b_k |x_1 t_k + x_2 - y_k|, \tag{6}$$

т. е. использовать не квадраты, а модули невязок.

Отыскание минимума функции (6) — задача линейного программирования. Для ее решения нельзя воспользоваться уравнениями вида (4), так как функция (6), в отличие от функции (5), недифференцируема. Преимущество метода наименьших квадратов в том, что вычисление обобщенного решения, понимаемого в смысле метода наименьших квадратов, существенно проще.

2. Задача уточнения результатов неточных измерений за счет числа этих измерений. Пусть заранее известно, что некоторая величина $y = y(t)$ зависит от своего аргумента t линейно, т. е. известно, что эта зависимость имеет вид (1). Задача состоит в том, чтобы по результатам измерения величины $y(t)$ при нескольких значениях t вычислить коэффициенты x_1, x_2. Допустим, что измерения произведены при $t = t_k$ $(k = 1, 2, 3, 4)$ и их результаты приведены на рис. 12.

Для определения x_1, x_2 снова получится та же переопределенная система (2). Эта система окажется несовместной, так как измерения производились с некоторой неизбежной погрешностью. Обобщенное

решение снова можно определить как пару чисел x_1, x_2, придающую наименьшее значение функции вида (5), и воспользоваться для этого системой вида (4).

Подчеркнем, что использование четырех неточных измерений и соответствующей переопределенной системы уравнений (2), вообще говоря, позволяет уменьшить влияние на результат погрешностей, допускаемых при каждом измерении.

Существует точная постановка в терминах теории вероятностей задачи о повышении надежности результатов, полученных при измерениях со случайными погрешностями за счет числа этих измерений и их обработки с помощью метода наименьших квадратов [9].

Переопределенные системы линейных уравнений могут возникать не только при построении эмпирических формул или при уточнении результатов неточных измерений за счет их числа (см. гл. 13, § 2).

§ 2. Переопределенные СЛАУ и обобщенные решения в общем случае

1. Каноническая запись переопределенной системы линейных алгебраических уравнений имеет следующий вид:

$$\begin{gathered} a_{11}x_1 + \ldots + a_{1s}x_s = f_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + \ldots + a_{ns}x_s = f_n, \\ n > s. \end{gathered} \tag{1}$$

Введем пространства R^s, R^n, состоящие из элементов вида $x = \begin{bmatrix} x_1 \\ \ldots \\ x_s \end{bmatrix}$, $f = \begin{bmatrix} f_1 \\ \ldots \\ f_n \end{bmatrix}$ и имеющие размерности s, n $(n > s)$ соответственно. Обозначим через A прямоугольную матрицу коэффициентов системы (1):

$$A = \begin{bmatrix} a_{11} & \ldots & a_{1s} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{ns} \end{bmatrix}. \tag{2}$$

Тогда систему (1) можно коротко записать в виде

$$Ax = f, \quad x \in R^s, \quad f \in R^n. \tag{3}$$

Введем в R^n “основное” скалярное умножение, положив

$$(f, g)^{(n)} = \sum_{k=1}^{n} f_k g_k. \tag{4}$$

Наряду с (4) можно ввести скалярное произведение в R^n множеством других способов. Именно, произвольной симметричной и положительно определенной матрице $B = B^* > 0$, т. е. $(Bf, f) > 0$ $(f \neq 0)$, соответствует скалярное умножение

$$[f, g]_B = (Bf, g), \quad f, g \in R^n. \tag{5}$$

Известно, что любое скалярное умножение в пространстве R^n можно задать формулой вида (5), подобрав подходящую матрицу $B = B^* > 0$.

Система (1), как правило, не имеет классического решения, т. е. не существует такого набора чисел $x_1, x_2, ..., x_s$, который обращает каждое из n уравнений (1) в тождество.

Определение. Фиксируем $B = B^* > 0$, $B\colon\ R^n \to R^n$. Введем функцию от $x \in R^n$, положив

$$\Phi(x) = [Ax - f, Ax - f]_B. \tag{6}$$

Примем за *обобщенное решение* x_B системы (1) тот вектор $x_B \in R^s$, который придает наименьшее значение квадратичной функции (6).

Замечание. Выбор $B = B^* > 0$ находится в руках исследователя. Матрица B имеет смысл "весовой" матрицы и выбирается из тех или иных соображений о том, какую цену придать невязке системы (1) при заданных $x_1, x_2, ..., x_s$.

Теорема 1. *Пусть столбцы матрицы A линейно независимы, т. е. ранг A равен s. Тогда существует одно и только одно обобщенное решение x_B системы* (1). *Обобщенное решение системы* (1) *является классическим решением системы уравнений*

$$A^*BAx = A^*Bf, \tag{7}$$

которая содержит s скалярных уравнений относительно s неизвестных $x_1, x_2, ..., x_s$.

Доказательство. Введем обозначение $\mathfrak{a}_k$ $(\mathfrak{a}_k \in R^n)$ для столбца с номером k $(k = 1, 2, ...)$ матрицы A, так что

$$\mathfrak{a}_k = \begin{bmatrix} a_{1k} \\ \dots \\ a_{nk} \end{bmatrix}, \quad k = 1, 2, ..., s. \tag{8}$$

Из формулы умножения матриц видим, что матрица $C = A^*BA$ системы (7) есть квадратная $s \times s$-матрица. Элемент c_{ij} этой матрицы, стоящий на пересечении i-й строки и j-го столбца, есть число

$$c_{ij} = (\mathfrak{a}_i, B\mathfrak{a}_j)^{(n)} = (B\mathfrak{a}_i, \mathfrak{a}_j)^{(n)} = [\mathfrak{a}_i, \mathfrak{a}_j]_B.$$

Из (8) видно, что $c_{ij} = c_{ji}$, т. е. $C = C^*$.

Покажем, что матрица C невырожденна и, более того, положительно определена:

$$(C\xi, \xi)^{(s)} > 0, \quad \xi \in R^s, \quad \xi \neq 0. \tag{9}$$

Напомним известную формулу

$$(f, A\xi)^{(n)} \equiv (A^* f, \xi)^{(s)}, \quad f \in R^n, \quad \xi \in R^s, \tag{10}$$

для проверки которой достаточно записать левую и правую части формулы (10) в развернутом виде.

Пусть $\xi \in R^s$, $\xi = \begin{bmatrix} \xi_1 \\ \dots \\ \xi_s \end{bmatrix} \neq \begin{bmatrix} 0 \\ \dots \\ 0 \end{bmatrix}$. Тогда

$$A\xi = \xi_1 \mathfrak{a}_1 + \xi_2 \mathfrak{a}_2 + \dots + \xi_s \mathfrak{a}_s$$

есть линейная комбинация системы линейно независимых векторов $\mathfrak{a}_k$, не все коэффициенты которой равны нулю. Поэтому $A\xi \neq \neq 0 \in R^n$. Но тогда скалярный квадрат вектора $A\xi$ положителен:

$$0 < [A\xi, A\xi]_B = (BA\xi, A\xi)^{(n)} = (A^* BA\xi, \xi) = (C\xi, \xi)^{(s)},$$

и (9) доказано. В силу невырожденности матрицы C система (7) имеет одно и только одно решение $x_B \in R^s$.

Покажем, что x_B является единственным обобщенным решением системы (7). Это значит, что для любого $x = x_B + \delta$ $(\delta \in R^s)$ справедливо строгое неравенство

$$\Phi(x_B + \delta) > \Phi(x_B), \quad \delta \neq 0. \tag{11}$$

Для доказательства этого неравенства предварительно заметим, что

$$[Ax_B - f, A\delta]_B = (B(Ax_B - f), A\delta)^{(n)} = (A^* BAx_B - A^* Bf, \delta)^{(s)} = 0. \tag{12}$$

Это равенство справедливо, поскольку $A^* BAx_B = A^* Bf$.

Докажем теперь (11), опираясь на (9), (12):

$$\Phi(x_B + \delta) = [A(x_B + \delta) - f, A(x_B + \delta) - f]_B =$$
$$= [Ax_B - f, Ax_B - f]_B - 2[Ax_B - f, A\delta]_B + [A\delta, A\delta]_B =$$
$$= \Phi(x_B) - 2[Ax_B - f, A\delta]_B + (C\delta, \delta)^{(s)} = \Phi(x_B) + (C\delta, \delta)^{(s)} > \Phi(x_B). \ \square$$

2. Замечания о вычислении обобщенного решения. Матрица системы (7) симметричная и положительно определенная. Этим можно воспользоваться при вычислении решения методом итераций. В случае если $\mathfrak{a}_1, \mathfrak{a}_2, \dots, \mathfrak{a}_s \in R^s$ — ортонормированная система векторов, т. е. если

$$[\mathfrak{a}_i, \mathfrak{a}_j]_B = \delta_{ij}, \quad i, j = 1, 2, \dots, s, \tag{13}$$

матрица C оказывается единичной, так как было показано, что $c_{ij} = = [a_i, a_j]_B$. В этом случае $x_B = A^* Bf$. Если условие (13) выполняется лишь приближенно, матрица C оказывается близка к единичной и

потому хорошо обусловлена, так что решение (7) можно легко вычислить итерациями.

3. Геометрический смысл метода наименьших квадратов. Систему (1) можно записать так:

$$x_1\mathfrak{a}_1 + x_2\mathfrak{a}_2 + \ldots + x_s\mathfrak{a}_s = f, \tag{14}$$

где $\mathfrak{a}_k = \begin{bmatrix} a_{1k} \\ \ldots \\ a_{nk} \end{bmatrix} \in R^n$, $f = \begin{bmatrix} f_1 \\ \ldots \\ f_n \end{bmatrix} \in R^n$. Требуется найти коэффициенты $x_1, x_2, \ldots, x_s$ линейной комбинации $x_1\mathfrak{a}_1 + x_2\mathfrak{a}_2 + \ldots + x_s\mathfrak{a}_s$ так, чтобы эта линейная комбинация наименее отличалась от f:

$$\left\| f - \sum_{k=1}^{s} x_k\mathfrak{a}_k \right\| \Rightarrow \min. \tag{15}$$

Обозначим через $R^{(s)}(\mathfrak{a}) \subset R^n$ подпространство размерности s пространства R^n, состоящее из всевозможных линейных комбинаций векторов $\mathfrak{a}_1, \mathfrak{a}_2, \ldots, \mathfrak{a}_s$. Если $x_1, x_2, \ldots, x_s$ — обобщенное решение системы (1), то линейная комбинация $\sum x_k\mathfrak{a}_k$ есть ортогональная в смысле скалярного умножения $[\,,\,]_B$ проекция вектора $f \in R^n$ на подпространство $R^s(\mathfrak{a})$.

В самом деле, любой вектор из $R^s(\mathfrak{a})$ имеет вид $A\delta = \delta_1\mathfrak{a}_1 + \delta_2\mathfrak{a}_2 + \ldots + \delta_s\mathfrak{a}_s \in R^s(\mathfrak{a})$ $(\delta \in R^s)$. Наименее уклоняется от f элемент $\sum x_k\mathfrak{a}_k$ подпространства $R^s(\mathfrak{a})$, имеющий вид Ax_B, где x_B — решение системы (7). Очевидно, что в силу (12) элемент $f - Ax_B$ ортогонален любому элементу $A\delta \in R^s(\mathfrak{a})$.

Если в пространстве $R^s(\mathfrak{a})$ вместо $\mathfrak{a}_1, \mathfrak{a}_2, \ldots, \mathfrak{a}_s$ выбрать какой-нибудь другой базис $\mathfrak{a}'_1, \mathfrak{a}'_2, \ldots, \mathfrak{a}'_s$, то система (7) заменится системой

$$C'x' = f \tag{16}$$

с матрицей $C'[c'_{ij}]$, $c'_{ij} = [A'\mathfrak{a}'_i, \mathfrak{a}'_j]_B$, $i, j = 1, 2, \ldots, s$, где A' — матрица, столбцы которой есть $\mathfrak{a}'_k$.

Вместо решения x_B системы (7) получим новое решение x'_B системы (16), но проекция f на $R^s(\mathfrak{a})$, очевидно, останется прежней, так что справедливо равенство

$$\sum x_k\mathfrak{a}_k = \sum x'_k\mathfrak{a}'_k.$$

Если нас интересует проекция заданного f на заданное подпространство $R^s \subset R^n$, то естественно стремиться к выбору базиса $\mathfrak{a}_1$, $\mathfrak{a}_2, \ldots, \mathfrak{a}_s$ этого подпространства, по возможности мало отличающегося от ортонормированного. Искомая проекция от выбора базиса в $R^s(\mathfrak{a})$ не зависит, а уравнение (7) в случае такого выбора базиса будет иметь

хорошо обусловленную, близкую к единичной матрицу (см. задачу 6 в конце параграфа).

4. Переопределенные системы, заданные в операторной форме. Наряду с канонической формой задания (1) переопределенная система может быть задана в операторной форме. Именно, пусть имеются два линейных пространства R^s, R^n размерностей s, n $(s < n)$. Пусть заданы линейный оператор $A\colon R^s \to R^n$ и вектор $f \in R^n$. Требуется найти $x \in R^s$ такой, чтобы выполнялись условия

$$Ax = f, \quad x \in R^s, \quad f \in R^n. \tag{17}$$

Эта запись по содержанию отличается от записи (3) тем, что здесь не предполагается, что x, f заданы своими координатами и что A задан матрицей.

Система (17), вообще говоря, не имеет классического решения, так как вектор $Ax \in R^n$ при любом x принадлежит s-мерному подпространству $R(A) \subset R^n$, в которое переходит R^s при преобразовании $A\colon R^s \to R^n$.

Для введения обобщенного решения надо определить в R^n скалярное умножение $(f,g)^{(n)}$. Все остальные скалярные умножения имеют вид

$$[f,g]_B = (Bf,g)^{(n)},$$

где $B = B^* > 0$. Введем функцию

$$\Phi_B(x) = [Ax - f, Ax - f]_B.$$

Примем за обобщенное решение x_B то значение $x \in R^s$, для которого эта функция принимает наименьшее значение:

$$\Phi_B(x) \Rightarrow \min.$$

Пример. Пусть в квадрате $(0 \leqslant x \leqslant 1,\, 0 \leqslant y \leqslant 1)$ введены две сетки — крупная с шагом $H = 0{,}1$ и с точками

$$(X_m, Y_n) = (mH, nH), \quad m,n = 0,1,...,10,$$

а также мелкая с шагом $h = 10^{-2}$ и с точками

$$(x_m, y_n) = (mh, nh), \quad m,n = 0,1,...,10^2.$$

Введем линейные пространства $U^{(H)}$, $F^{(h)}$. К первому отнесем все функции u_{mn}, определенные на крупной сетке, а ко второму — все функции f_{mn}, определенные на мелкой сетке. Зададим оператор $A\colon U^{(H)} \to F^{(h)}$, который по функции u_{mn} строит функцию f_{mn}, доопределяя u_{mn} в точках мелкой сетки с помощью линейной интерполяции по каждому из аргументов x, y. Пусть путем приближенных измерений некоторого физического скалярного поля в точках мелкой сетки получена некоторая функция $f \in F^{(h)}$, $f = \{f_{mn}\}$.

Поставим задачу — отыскать такую функцию $u \in U^{(H)}$, чтобы она была обобщенным решением следующей переопределенной системы:

$$Au = f, \quad u \in U^{(H)}, \quad f \in F^{(h)}. \tag{18}$$

Обобщенное решение будем понимать в смысле следующего скалярного умножения в $F^{(h)}$:

$$(f, g) = \sum_{m,n=0}^{100} f_{mn} g_{mn}, \quad f, g \in F^{(h)}.$$

Содержательный смысл этой задачи — произвести сглаживание экспериментальных данных, уменьшив влияние случайных погрешностей, допущенных при измерениях. В результате этого сглаживания вместо экспериментальной функции $f \in F^{(h)}$ можно будет использовать функцию $\widetilde{f} = Au$.

Мы не будем приводить здесь алгоритм отыскания обобщенного решения $u \in U^{(H)}$ задачи (18), отнеся это к числу задач для самостоятельного решения. Подчеркнем только, что задача (18) поставлена не в канонической форме (1), а в операторной.

Можно многими способами перейти от операторной формы (18) к канонической форме (1) задания переопределенной системы (как это сделать?), однако это нецелесообразно.

Задачи

1. Найти обобщенное решение (в смысле метода наименьших квадратов) переопределенной системы

$$x + y = 1, \quad x - y = 2, \quad 2x + y = 2{,}4.$$

2. Произведено некоторое число m приближенных измерений длины l. Получились следующие результаты: $l = l_1, l = l_2, ..., l = l_m$, где l_j — некоторые числа.

Найти решение в смысле метода наименьших квадратов этой системы m уравнений относительно одного неизвестного l.

3. Электрическое сопротивление проволоки R линейно зависит от температуры t, так что $R = a + bt$.

Определить a, b по результатам следующих неточных измерений:

t	19,1	25,0	30,1	36,0
R	76,30	77,80	79,75	80,80

Найти сопротивление проволоки при $t = 21{,}28$.

4. Скорость v корабля связана с мощностью p двигателя эмпирической формулой $p = a + bv^3$.

Определить a, b по данным измерений, приведенным в таблице

v	6	8	10	12
r	420	805	1370	2370

5. Измерения углов α, β, γ треугольника на плоскости привели к значениям $\alpha = 52^\circ 5'$, $\beta = 50^\circ 1'$, $\gamma = 78^\circ 6'$. Сумма результатов измерений есть $180^\circ 12'$ и дает невязку $12'$, происходящую от погрешностей измерения.

Ликвидировать невязку, следуя предписаниям метода наименьших квадратов.

6. Некоторая функция задана таблицей своих значений $y_k = 1 + x_k + x_k^2 + \sin x_k$ на множестве точек

$$x_k = -1 + 10^{-2}k, \quad k = 0, 1, ..., 200.$$

Требуется с помощью компьютера построить многочлен $Q_m(x)$ степени не выше заданного $m \leqslant 50$, для которого сумма

$$\sum_{k=0}^{200} |Q_m(x_k) - y_k|^2$$

принимает наименьшее значение.

Рассмотреть и сравнить следующие три варианта решения и вычислить значения Q_m в точках

$$x_{k+1/2} = \frac{x_k + x_{k+1}}{2}, \quad k = 0, 1, ..., 199.$$

Вариант 1. Ищем многочлен $Q_m(x)$ в виде $Q_m(x) = \sum\limits_{s=0}^{m} c_s x^s$ с неопределенными коэффициентами c_s.

Вариант 2. Ищем многочлен $Q_m(x)$ в виде

$$Q_m(x) = \sum_{s=0}^{m} c_s P_s(x),$$

где $P_s(x)$ — многочлен Лежандра степени s. О многочленах Лежандра известно, что

$$P_0(x) = 1, \quad P_1(x) = x,$$

$$P_{l+1}(x) = \frac{2l}{l+1}\, x P_l(x) - \frac{l}{l+1}\, P_{l-1}(x), \quad l = 1, 2, ...$$

Они ортогональны на отрезке $-1 \leqslant x \leqslant 1$, т. е.

$$\int_{-1}^{1} P_k(x) P_l(x)\, dx = 0, \quad k \neq l.$$

Далее,

$$\int_{-1}^{1} P_k^2(x)\, dx = \frac{2}{2k+1}.$$

Вариант 3. Ищем многочлен $Q_m(x)$ в виде

$$Q_m(x) = \sum_{s=1}^{m} c_s \widetilde{P}_s(x),$$

где

$$\widetilde{P}_s(x) = \sqrt{\frac{2s+1}{2}}\, P_s(x).$$

Построение алгоритма отыскания коэффициентов c_s $(s = 0, 1, ..., m)$ во всех трех вариантах осуществить по следующему плану.

1°. Вычислить матрицу A^*BA и матрицу A^*B, с помощью которых записывается система

$$A^*BAc = A^*By, \quad c = \begin{bmatrix} c_0 \\ ... \\ c_m \end{bmatrix}, \quad y = \begin{bmatrix} y_0 \\ ... \\ y_{200} \end{bmatrix}$$

метода наименьших квадратов.

2°. Написать программу решения полученной системы методом сопряженных градиентов, принимая за нулевое приближение $\overline{c}^{(0)} = \begin{bmatrix} 0 \\ \cdots \\ 0 \end{bmatrix}$ и заканчивая итерации, если $c^{(p)}$ отличается от $c^{(p-1)}$ “достаточно мало”, а именно

$$\max_s | c_s^{(p)} - c_s^{(p-1)} | \leqslant \varepsilon, \quad \varepsilon = 10^{-3}.$$

3°. Объяснить, почему система $A^*BAc = A^*By$, которая возникает в варианте 1, обусловлена плохо, а в варианте 3 ее число обусловленности не сильно превосходит единицу.

7*. Составить программу для численного решения переопределенной системы, заданной в операторной форме в примере, который приведен в конце параграфа.

Указание. Ввести скалярное умножение в пространстве $U^{(H)}$, $F^{(h)}$, положив

$$(f, g)^{(h)} = h^2 \sum_{m,n=0}^{10} f_{mn} g_{mn}, \quad f, g \in F^{(h)}.$$

Реализовать на ЭВМ программу решения итерациями системы линейных уравнений $A^*Au = A^*f$, заданной в операторной форме.

ГЛАВА 7

ЧИСЛЕННОЕ РЕШЕНИЕ НЕЛИНЕЙНЫХ СКАЛЯРНЫХ УРАВНЕНИЙ И СИСТЕМ УРАВНЕНИЙ

Здесь мы изложим основные способы отыскания решений нелинейных скалярных уравнений

$$F(x) = 0, \tag{1}$$

где $F(x)$ — некоторая заданная функция, например $F(x) = \sin x - 0{,}5x$ или $F(x) = e^{-x} + \cos x$.

Будем рассматривать также системы скалярных уравнений

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}, \tag{2}$$

где $\mathbf{F}(\mathbf{x})$ — некоторая заданная вектор-функция от векторного аргумента $\mathbf{x} = (x_1, x_2, ..., x_n)$, $x_1, x_2, ..., x_n$ — скалярные координаты вектора x; например, в случае

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} F_1(x_1, x_2) \\ F_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} x_1^2 + x_2^2 - 25 \\ x_2 - x_1^2 \end{bmatrix}$$

система (2) в развернутом виде запишется так:

$$\begin{aligned} x_1^2 + x_2^2 - 25 = 0, \\ x_2 - x_1^2 = 0. \end{aligned} \qquad (3)$$

При решении скалярных уравнений $F(x) = 0$, а также систем $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ таких уравнений возникают две задачи: задача указания областей, в которых лежит по одному решению (задача отделения корней), и задача отыскания корней с заданной точностью (задача уточнения корней).

Для отделения корней нет общих приемов. Используются графики, участки монотонности функции, на которых она меняет знак, и другие частные приемы. Имеется лишь один большой класс функций — многочлены с вещественными коэффициентами, для которого задача отделения вещественных корней решена в общем виде и полностью. Такое решение дает теорема Штурма.

Для уточнения корней используются различные варианты последовательных приближений (итераций).

§ 1. Метод простых итераций

Рассмотрим сначала случай одного уравнения $F(x) = 0$, а затем случай системы $\mathbf{F}(\mathbf{x}) = \mathbf{0}$.

1. Случай одного скалярного уравнения. Пусть известно, что интересующий нас корень x уравнения

$$F(x) = 0 \qquad (1)$$

лежит на интервале $a < x < b$. Каким-нибудь способом приводим уравнение (1) к равносильному ему на интервале $U = \{a < x < b\}$ уравнению вида $x = f(x)$. Можно положить

$$f(x) = x - \alpha(x)F(x), \qquad (2)$$

где $\alpha(x)$ — произвольная, не обращающаяся в нуль в точках U функция. Для отыскания решения $\check{x}$, принадлежащего интервалу $a < x < b$, зададим x_0, а затем последовательно вычислим $x_1, x_2, ...$ по формуле

$$x_{p+1} = f(x_p), \quad p = 0, 1, ... \qquad (3)$$

При этом предполагается, что каждое x_p принадлежит области определения функции $f(x)$, так что последовательное вычисление чисел $x_1, x_2, ..., x_p, ...$ возможно.

Теорема 1. *Пусть $f(x)$ непрерывна в U, и пусть при заданном x_0 последовательность* (3) *сходится к некоторому $\check{x}$, принадлежащему U. Тогда $\check{x}$ — корень уравнения* (2), *т. е. выполнено равенство $\check{x} = f(\check{x})$.*

Доказательство. Очевидно, что $\lim\limits_{n\to\infty} x_{n+1} = \lim\limits_{n\to\infty} f(x_n)$. Но в силу определения непрерывности $\lim\limits_{n\to\infty} f(x_n) = f\big(\lim\limits_{n\to\infty} x_n\big)$. Поэтому $\lim\limits_{n\to\infty} x_{n+1} = f\big(\lim\limits_{n\to\infty} x_n\big)$, или $\check{x} = f(\check{x})$. □

О факте сходимости практически можно судить, наблюдая за последовательностью x_n в процессе вычисления ее членов на компьютере, "поручив" установление этого факта самому компьютеру. Если сходимость установлена, то за корень $\check{x}$ можно приближенно принять член $x^{(p)}$ последовательности (3) с достаточно большим номером p.

Однако последовательность (3) не всегда сходится. На рис. 13

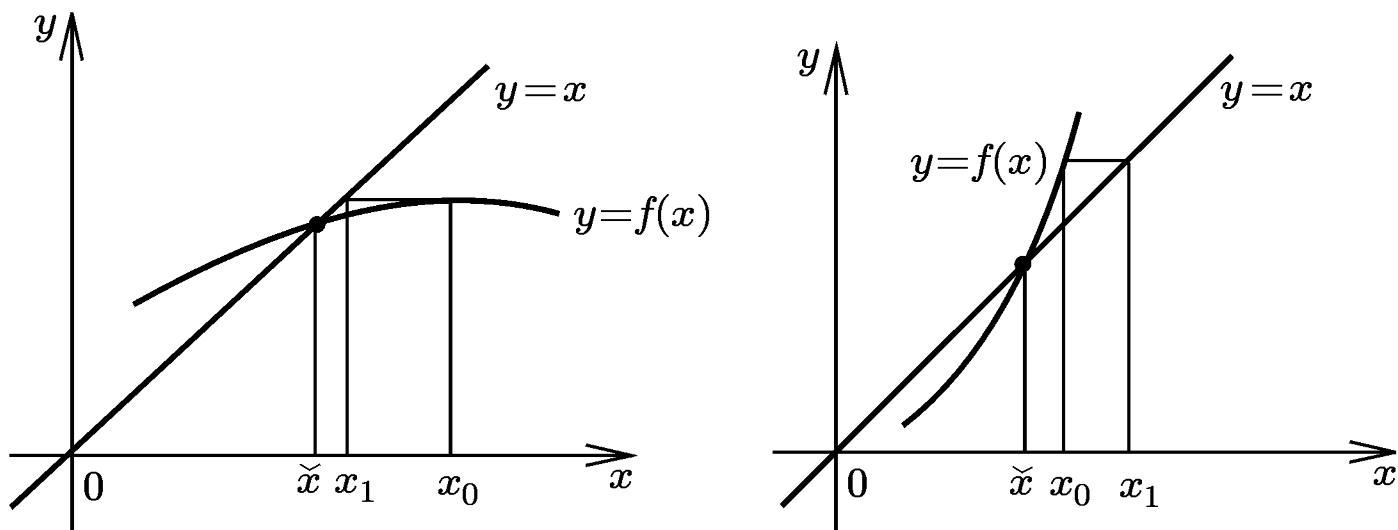


Рис. 13

изображены графики $f(x)$, для одного из которых при указанном на рисунке нулевом приближении x_0 последовательность (3) сходится, а для другого — нет.

Сформулируем условия, при которых гарантирована сходимость (3). Введем для этого следующее определение. Отображение (функция) $f(x)$ называется *сжимающим* в области U с коэффициентом сжатия q $(0 \leqslant q < 1)$, если для любых двух x', x'' из U выполнено неравенство

$$|f(x'') - f(x')| \leqslant q\,|x'' - x'|. \tag{4}$$

Теорема 2. *Пусть уравнение $x = f(x)$ имеет решение $\check{x}$, принадлежащее области U, и пусть отображение $f(x)$ является сжимающим в области U с некоторым коэффициентом сжатия q $(0 \leqslant q < 1)$. Тогда решение $\check{x}$ является единственным решением в области U и существует столь малое $R > 0$, что при выборе x_0 из условия $|\check{x} - x_0| < R$ все члены последовательности* (3) *будут определены, причем вы-*

полняются неравенства

$$|\check{x} - x_n| \leqslant q^n |\check{x} - x_0|, \quad n = 0, 1, ..., \tag{5}$$

т. е. имеет место сходимость x_n к $\check{x}$ со скоростью геометрической прогрессии со знаменателем q.

Доказательство. Сначала докажем единственность решения. Допустим, что наряду с $\check{x}$ существует решение $\check{x}'$:

$$\check{x}' = f(\check{x}'), \quad \check{x}, \check{x}' \in U.$$

Тогда в силу того, что $f(x)$ — сжимающее отображение, $|f(\check{x}) - f(\check{x}')| \leqslant |q(\check{x} - \check{x}')|$. Но в силу $f(\check{x}) = \check{x}$ и $f(\check{x}') = \check{x}'$ отсюда получаем

$$|\check{x}' - \check{x}| < |\check{x}' - \check{x}|.$$

Противоречие доказывает единственность.

Далее возьмем $R > 0$ столь малым, чтобы совокупность $U(\check{x}, R)$ точек, удовлетворяющих неравенству $|x - \check{x}| < R$, целиком принадлежала U. Пусть $x_0 \in U(\check{x}, R)$. Тогда

$$|\check{x} - x_1| = |f(\check{x}) - f(x_0)| \leqslant q|\check{x} - x^0| \leqslant qR < R,$$

так что x_1 тоже принадлежит окрестности $U(\check{x}, R)$. Аналогично x_2, $x_3, ..., x_n, ...$ лежат в этой окрестности.

При $n = 0$ оценка (5) тривиальна. Допустим, она уже доказана для $n = 0, 1, ..., k$. Докажем ее для $n = k + 1$:

$$|\check{x} - x_{k+1}| = |f(\check{x}) - f(x_k)| \leqslant q|\check{x} - x_k| \leqslant qq^k|\check{x} - x_0|.$$

Таким образом, неравенство (5) доказано для всех n по индукции. □

На первый взгляд доказанная теорема ничего не дает для отыскания $\check{x}$, поскольку мы не можем указать фактически то $R > 0$ и ту окрестность $U(\check{x}, R)$, в которой надо выбирать x_0. Однако сам факт существования такой окрестности позволяет организовать автоматическое вычисление $\check{x}$ на ЭВМ. Именно, зададим произвольно $x_0 \in U$. Если последовательность (3) удается вычислить и она сходится к некоторому $\check{x}$, то в силу теоремы 1 это и есть искомый корень. Если сходимости нет, то программа расчета должна взять другое значение x_0' и повторить расчет; если сходимости опять нет, то выбираем за начальное приближение третью точку x_0'' и т. д. Если в процессе выработки этих начальных приближений $x_0, x_0', x_0'', ...$ они все "гуще" располагаются в U, то при выборе x_0 на некотором шаге мы неизбежно попадем в достаточно малую окрестность $U(\check{x}, R)$ корня, и итерационный процесс сойдется.

Укажем достаточный признак того, что отображение является сжимающим.

Теорема 3. *Пусть $f(x)$ имеет производную во всех точках области U (т. е. интервала $a < x < b$), и пусть существует q $(0 \leqslant$*

$\leqslant q < 1$, $q = \mathrm{const}$) *такое, что* $|f'(x)| \leqslant q$ *всюду в* U. *Тогда отображение* $f(x)$ *сжимающее в* U *с коэффициентом сжатия* q.

Доказательство. По теореме о конечных приращениях Лагранжа для любых x', x'' из U

$$f(x'') - f(x') = (x'' - x')f'(\xi),$$

где ξ — некоторая точка, лежащая между x' и x''. Но тогда

$$|f(x'') - f(x')| = |x'' - x'||f'(\xi)| \leqslant q|x'' - x'|. \square$$

2. Случай системы уравнений. Пусть задана система уравнений

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}, \tag{6}$$

причем вектор-функция $\mathbf{F}(\mathbf{x})$ определена в некоторой области U пространства R^n, $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, и имеет решение $\check{x}$ в этой области. Приведем уравнение (6) к равносильному уравнению вида

$$\mathbf{x} = \mathbf{f}(\mathbf{x}). \tag{7}$$

Это, очевидно, можно сделать многими способами, например, положив $\mathbf{f}(\mathbf{x}) = \mathbf{x} - \alpha(\mathbf{x})\mathbf{F}(\mathbf{x})$, где $\alpha(\mathbf{x})$ — произвольная матричная функция, определитель которой отличен от нуля всюду в U.

Зададим $x_0 \in U$ и построим последовательность

$$\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n), \quad n = 0, 1, ... \tag{8}$$

Будем считать пространство R^n нормированным. Тогда остаются справедливыми все определения и утверждения теоремы 1, относящиеся к случаю скалярного уравнения. Нужно только всюду вместо абсолютной величины (чисел) использовать нормы (векторов).

Признак того, что отображение $\mathbf{f}\colon R^n \to R^n$ является сжимающим, сформулированный в теореме 2 для скалярной функции, т. е. в случае $n = 1$, в общем случае формулируется несколько сложнее.

Область U пространства R^n называется *выпуклой*, если наряду с любыми двумя точками $\mathbf{a} = (a_1, a_2, ..., a_n)$, $\mathbf{b} = (b_1, b_2, ..., b_n)$, принадлежащими области U, ей принадлежат и все точки отрезка $\mathbf{ab}$, содержащего эти точки, т. е. все точки вида

$$\mathbf{x} = \mathbf{a} + t(\mathbf{b} - \mathbf{a}), \tag{9}$$

где t изменяется от 0 до 1.

Теорема 4. *Пусть область* $U \subset R^n$ *выпуклая, и пусть компоненты* $f_i(x_1, x_2, ..., x_n)$ *вектор-функции*

$$\mathbf{f}(\mathbf{x}) = \begin{cases} f_1(x_1, x_2, ..., x_n), \\ \dots\dots\dots\dots\dots \\ f_n(x_1, x_2, ..., x_n) \end{cases}$$

имеют равномерно непрерывные производные первого порядка в U. *Рассмотрим матрицу*

$$\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \partial f_1/\partial x_1 & \ldots & \partial f_1/\partial x_n \\ \ldots & \ldots & \ldots \\ \partial f_n/\partial x_1 & \ldots & \partial f_n/\partial x_n \end{bmatrix}.$$

Если норма этой матрицы при всех $\mathbf{x} \in U$ *не превосходит некоторого числа* q $(0 \leqslant q < 1)$, *то отображение* $\mathbf{f}(\mathbf{x})$ *является сжимающим в области* U:

$$\|\mathbf{f}(\mathbf{x}'') - \mathbf{f}(\mathbf{x}')\| \leqslant q\,\|\mathbf{x}'' - \mathbf{x}'\|,$$

где $\mathbf{x}'$, $\mathbf{x}''$ — *произвольные точки из* U.

Доказательства этой теоремы из математического анализа мы не приводим.

Для вычисления решения $\check{\mathbf{x}}$ уравнения $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ методом простых итераций (т. е. с помощью равносильного уравнения $\mathbf{x} = \mathbf{f}(\mathbf{x})$ и последовательности $\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n)$, $n = 0, 1, \ldots$; x_0 задано) надо стараться выбрать это равносильное уравнение так, чтобы отображение f в области U, содержащей решение $\check{\mathbf{x}}$, было сжимающим с коэффициентом сжатия q, как можно более близким к нулю.

§ 2. Метод линеаризации Ньютона

Сначала рассмотрим случай одного уравнения $F(x) = 0$, а потом — системы $\mathbf{F}(\mathbf{x}) = \mathbf{0}$.

1. Метод линеаризации для одного уравнения. Зададим x_0. Пусть некоторое приближение x_n к корню $\check{x}$ уравнения $F(x) = 0$ уже найдено. Воспользуемся приближенной формулой

$$F(x) \approx F(x_n) + F'(x_n)(x - x_n),$$

точность которой возрастает при приближении x к x_n. Вместо исходного уравнения $F(x) = 0$ воспользуемся линейным уравнением

$$F(x_n) + F'(x_n)(x - x_n) = 0.$$

Решение этого уравнения примем за приближение x_{n+1}:

$$x_{n+1} = x_n - [F'(x_n)]^{-1} F(x_n), \quad n = 0, 1, \ldots \tag{1}$$

Метод линеаризации Ньютона допускает простую геометричес-

кую интерпретацию (рис. 14). Криволинейный график функции $F(x)$ заменяется касательной к нему в точке $(x_p, F(x_p))$. За приближение x_{n+1} принимается точка пересечения этой касательной с осью абсцисс.

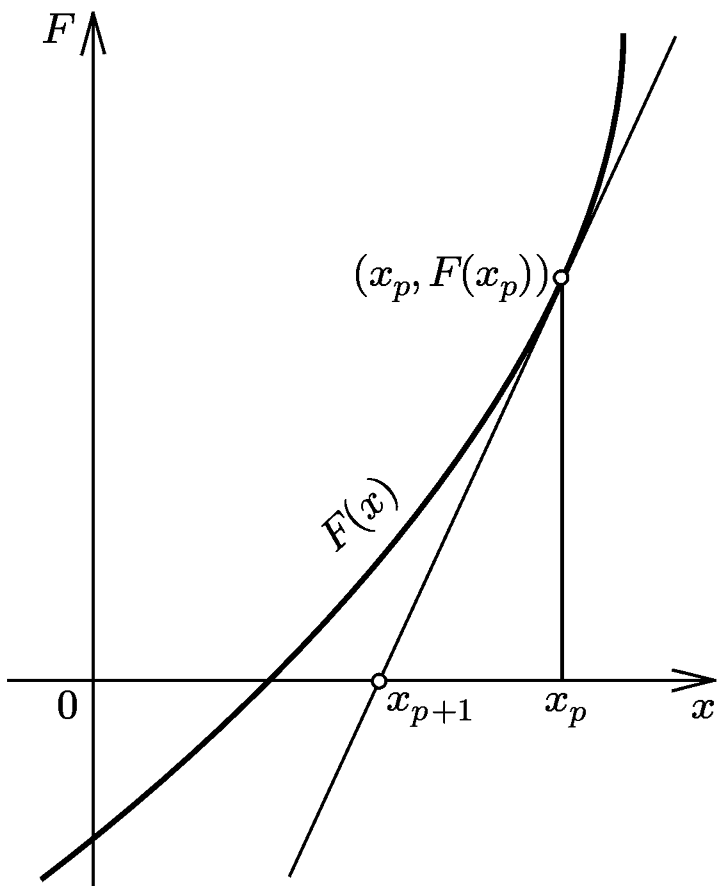


Рис. 14

Формулу (1) можно интерпретировать следующим образом. Считая, что $F'(x) \neq 0$, переходим от уравнения $F(x) = 0$ к равносильному уравнению

$$x = f(x),$$
$$f(x) = x - [F'(x)]^{-1} F(x)$$

и пользуемся простыми итерациями. Очевидно, что в точке $\check{x}$, где $F(\check{x}) = 0$,

$$df/dx|_{x=\check{x}} = 0.$$

Будем считать, что $f''(x)$ непрерывна. Тогда для каждого $0 < q < 1$ найдется (малая) окрестность точки $\check{x}$, в которой $|f'(x)| \leqslant q$.

В силу теоремы 3 из § 1 отображение $f(x)$ в этой окрестности будет сжимающим с этим коэффициентом сжатия q, т. е. тем более сжимающим, чем меньше окрестность корня $\check{x}$.

В соответствии с этим последовательность (1) (т. е. приближения, полученные методом линеаризации) будет сходиться к $\check{x}$, если точку x_0 выбрать достаточно близко к $\check{x}$, причем скорость убывания погрешности $|\check{x} - x_p|$ будет быстрее, чем скорость убывания членов геометрической прогрессии q^p, каково бы ни было фиксированное q $(0 < q < 1)$.

2. Метод линеаризации для системы. Рассмотрим систему

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}. \qquad (2)$$

Пусть в некоторой области U есть решение $\check{\mathbf{x}}$ системы (2). Будем считать, что компоненты $F_j(x_1, x_2, ..., x_n)$ $(j = 1, 2, ...)$ функций, составляющих вектор-функцию $\mathbf{F}(\mathbf{x})$, являются достаточно гладкими функциями.

Пусть приближение $\mathbf{x}_n$ к решению уже найдено. Воспользуемся приближенной формулой

$$\mathbf{F}(\mathbf{x}) \approx \mathbf{F}(\mathbf{x}_n) + \left.\frac{d\mathbf{F}(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_n} (\mathbf{x} - \mathbf{x}_n)$$

и вместо уравнения (2) рассмотрим линейную систему уравнений

$$\mathbf{F}(\mathbf{x}_p) + \left.\frac{d\mathbf{F}(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_p} (\mathbf{x} - \mathbf{x}_n) = 0. \tag{3}$$

Решение этой системы примем за $\mathbf{x}_{p+1}$.

Очевидно, что $\mathbf{x}_{p+1}$ выражается формулой

$$\mathbf{x}_{p+1} = \mathbf{x}_p - \left[\frac{d\mathbf{F}(\mathbf{x}_p)}{d\mathbf{x}}\right]^{-1} \mathbf{F}(\mathbf{x}_p), \tag{4}$$

однако фактически решение линейной системы (3) можно искать одним из изложенных выше способов решения систем линейных уравнений (методы Гаусса и другие).

Для реализуемости метода Ньютона надо предполагать, что матрица $d\mathbf{F}(\mathbf{x})/d\mathbf{x}$ в точке $\mathbf{x} = \check{\mathbf{x}}$ невырожденна, так что $[d\mathbf{F}(\mathbf{x})/d\mathbf{x}]^{-1}$ существует. В силу непрерывности она невырожденна и в малой окрестности точки $\check{\mathbf{x}}$.

Как и в случае одного уравнения, можно считать, что мы перешли от (2) к равносильному уравнению

$$\mathbf{x} = \mathbf{f}(\mathbf{x}),$$

причем

$$\mathbf{f}(\mathbf{x}) = \mathbf{x} - \left[\frac{d\mathbf{F}(\mathbf{x})}{d\mathbf{x}}\right]^{-1} \mathbf{F}(\mathbf{x}), \tag{5}$$

а затем воспользовались методом простых итераций. Можно проверить, что матрица $d\mathbf{f}(\mathbf{x})/d\mathbf{x}$ в точке $\check{\mathbf{x}}$ есть нулевая матрица. Поэтому, как и в случае одного уравнения, отображение $\mathbf{f}(\mathbf{x})$, задаваемое формулой (5), является сжимающим в окрестности точки $\check{\mathbf{x}}$ с коэффициентом сжатия q, который тем ближе к нулю, чем меньше эта окрестность.

Доказательство аналогично случаю одного уравнения, но вместо теоремы 2 из § 1 опирается на теорему 3 из § 1 и более громоздко. В силу теоремы 4 из § 1 при выборе $\mathbf{x}_0$ из достаточно малой окрестности решения $\check{\mathbf{x}}$ процесс последовательных приближений (4) сходится и скорость убывания погрешности быстрее геометрической прогрессии q^p, как бы ни было мало фиксированное q $(0 < q < 1)$.

Задачи

1*. Пусть уравнение $F(x) = 0$ имеет корень $\check{x}$ на отрезке $[a, b]$, причем на этом отрезке функция имеет ограниченную вторую производную, а $F'(x) > 0$.

Доказать, что погрешности $|\check{x} - x_p|$ приближений x_p, полученные методом Ньютона, удовлетворяют оценкам

$$|\check{x} - x_{p+1}| \leqslant \text{const}|\check{x} - x_p|^2, \quad p = 0, 1, ...,$$

если только нулевое приближение x_0 выбрано достаточно близко к $\check{x}$.

2. Указать алгоритм для вычисления $\sqrt{5}$ с заданным числом верных десятичных знаков, рассматривая $\sqrt{5}$ как решение уравнения $x^2 - 5 = 0$. Воспользоваться методом Ньютона. Показать, что метод Ньютона сходится при произвольном выборе $x_0 > 0$.

3. Пусть графики некоторых заданных дифференцируемых функций $y = \varphi(x)$, $y = \psi(x)$ построены на плоскости Oxy и пересекаются в некоторой точке $x = \check{x}$. Для отыскания корня $\check{x}$ уравнения $F(x) \equiv \varphi(x) - \psi(x) = 0$ предполагается воспользоваться методом Ньютона. Пусть приближение x_p уже найдено и отмечено на оси абсцисс.

а) Построить следующее приближение x_{p+1} метода Ньютона, указав точку x_{p+1} на оси абсцисс.

б) Пусть графики $y = \varphi(x)$, $y = \psi(x)$ направлены один выпуклостью вверх, а другой — выпуклостью вниз. Показать, что тогда метод Ньютона сходится для любого $x_0 > \check{x}$.

4. Указать алгоритмы для вычисления вещественных решений следующих систем скалярных уравнений и вычислить решения, получив результаты с пятью верными знаками:

а) $\sin x - y = 1{,}30, \quad \cos y - x = -0{,}84;$

б) $x^2 + 4y^2 = 1, \quad x^4 + y^4 = 0{,}5.$

5. Для численного решения краевой задачи

$$\frac{d^2y}{dx^2} - y^3 = x^2, \quad 0 \leqslant x \leqslant 1, \quad y(0) = 1, \quad y(1) = 3,$$

на отрезке $0 \leqslant x \leqslant 1$ введена сетка $x_k = k/N$ $(k = 0, 1, \ldots, N)$, а для неизвестных $y_k = y(x_k)$ составлена разностная схема

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - y_k^3 = (kh)^2, \quad k = 1, 2, \ldots, N-1,$$

$$y_0 = 1, \quad y_N = 3, \quad h = 1/N.$$

Для решения этой системы $(N+1)$-го нелинейного уравнения относительно того же числа неизвестных $y_0, y_1, \ldots, y_N$ воспользоваться методом Ньютона.

а) Примем за начальное приближение $y^{(0)} = \{y_0^{(0)}, y_1^{(0)}, \ldots, y_N^{(0)}\}$ функцию $y_k^{(0)} = 1 + 2(kh)$ $(k = 0, 1, \ldots, N)$, определенную на сетке x_k и удовлетворяющую поставленным краевым условиям $y_0 = 1, y_N = 3$. Построить линейную систему метода Ньютона для отыскания по уже найденному приближению $y^{(p)} = \{y_k^{(p)}\}$ поправки $\varepsilon^{(p)} = \{\varepsilon_k^{(p)}\}$ для получения следующего приближения:

$$y^{(p+1)} = \{y_k^{(p)} + \varepsilon_k^{(p)}\}.$$

б) Показать, что линейную систему уравнений для поправки $\varepsilon^{(p)}$

$$\frac{\varepsilon_{k+1}^{(p)} - 2\varepsilon_k^{(p)} + \varepsilon_{k-1}^{(p)}}{h^2} - 3(y_k^{(p)})^2\varepsilon_k^{(p)} = (kh)^2 + (y_k^{(p)})^3, \qquad k = 1, 2, \ldots, N-1,$$

$$\varepsilon_0^{(p)} = \varepsilon_N^{(p)} = 0,$$

можно решить методом прогонки.

в) Составить программу для вычисления методом Ньютона решения исходной краевой задачи на компьютере и найти таблицу значений решения, подобрав с помощью экспериментальных расчетов число N и число итераций так, чтобы таблица $y^{(p)} = \{y_k^{(p)}\}$ имела четыре верных десятичных знака.

ЧАСТЬ III

МЕТОД КОНЕЧНЫХ РАЗНОСТЕЙ ДЛЯ ЧИСЛЕННОГО РЕШЕНИЯ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Самыми распространенными методами вычисления решений задач для обыкновенных дифференциальных уравнений и уравнений с частными производными являются метод конечных разностей и его модификации. Во всех вариантах этого метода в области определения искомых функций вводится сетка и решение ищется на сетке. Для значений искомой сеточной функции строится система скалярных уравнений, решение которой и служит приближенной таблицей значений решения исходной задачи.

Простейший способ построения этой системы скалярных уравнений — разностной схемы — состоит в приближенной замене производных, входящих в дифференциальное уравнение и в краевые условия, разностными отношениями, чем и объясняется название метода.

ГЛАВА 8

ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Пусть на некотором отрезке $a \leqslant x \leqslant b$ задано обыкновенное дифференциальное уравнение (или система уравнений) и требуется найти решение этого уравнения, удовлетворяющее дополнительным начальным или краевым условиям.

§ 1. Примеры разностных схем. Сходимость

Будем записывать краевые задачи символическим равенством

$$Lu = f. \tag{1}$$

Приведем примеры таких задач.

Пример задачи с начальными условиями для обыкновенного диф-

ференциального уравнения первого порядка:

$$\frac{du}{dx} + \frac{x}{1+u^2} = \cos x, \quad 0 \leqslant x \leqslant 1, \quad u(0) = 3. \tag{2}$$

Пример задачи с начальными условиями для дифференциального уравнения второго порядка:

$$\begin{gathered} \frac{d^2u}{dx^2} = (1+x^2)\,u + \sqrt{x}, \quad 0 \leqslant x \leqslant 1, \\ u(0) = 2, \quad \left.\frac{du}{dx}\right|_{x=0} = 1. \end{gathered} \tag{3}$$

Пример задачи с краевыми условиями, поставленными в начале $x = 0$ и в конце $x = 1$ отрезка:

$$\begin{gathered} \frac{d^2u}{dx^2} = (1+x^2)u + \sqrt{x}, \quad 0 \leqslant x \leqslant 1, \\ u(0) = 2, \quad u(1) = 1. \end{gathered} \tag{4}$$

Пример задачи с начальными условиями для системы двух уравнений первого порядка:

$$\begin{gathered} \frac{dv}{dx} + xvw = x^2 - 3x + 1, \quad 0 \leqslant x \leqslant 1, \\ \frac{dw}{dx} + \frac{1}{1+x^2}(v+w) = \cos^2 x, \quad 0 \leqslant x \leqslant 1, \\ v(0) = 1, \quad w(0) = -3. \end{gathered} \tag{5}$$

Решением этой задачи является вектор-функция $u = (v, w)$, компоненты которой удовлетворяют равенствам (5).

Во всех примерах мы рассматриваем задачу на отрезке $\overline{D} = \{0 \leqslant \leqslant x \leqslant 1\}$, а не на каком-либо другом только для определенности.

1. Примеры разностных схем. Будем предполагать, что решение $u(x)$ задачи (1) на отрезке $0 \leqslant x \leqslant 1$ существует. Для вычисления этого решения с помощью метода конечных разностей (или метода сеток) надо прежде всего выбрать на отрезке $\overline{D}$ конечное число точек, совокупность которых будем называть *сеткой* и обозначать через D_h, а затем считать искомым не решение $u(x)$ задачи (1), а таблицу $[u]_h$ значений этого решения в точках сетки D_h. Предполагается, что сетка D_h зависит от параметра h, который может принимать сколь угодно малые положительные значения. При стремлении шага сетки h к нулю сетка должна становиться все гуще. Например, можно положить $h = 1/N$, где N — какое-нибудь натуральное число, и принять за сетку D_h совокупность точек $x_0 = 0$, $x_1 = h$, $x_2 = 2h$, ..., $x_N = 1$. Искомая сеточная функция $[u]_h$ в этом случае в точке x_n сетки принимает значение, которое будем обозначать u_n.

Для приближенного вычисления таблицы значений решения $[u]_h$ в случае задачи (2) можно воспользоваться, например, системой уравнений

$$\frac{u_{n+1} - u_n}{h} + \frac{x_n}{1+u_n^2} = \cos x_n, \quad n = 0, 1, ..., N-1, \quad u_0 = 3, \tag{6}$$

полученной в результате замены производной du/dx в точках сетки разностными отношениями по приближенной формуле

$$\frac{du}{dx} \approx \frac{u(x+h)-u(x)}{h}.$$

Решение $u^{(h)} = (u_0^{(h)}, u_1^{(h)}, ..., u_N^{(h)})$ системы (6) определено на той же сетке D_h, что и искомая сеточная функция $[u]_h$. Его значения $u_0^{(h)}$, $u_1^{(h)}, ..., u_N^{(h)}$ в точках $x_0, x_1, ..., x_N$ последовательно вычисляются из системы (6) при $n = 0, 1, ..., N-1$. Для краткости в уравнениях (6) мы не пишем значок h при $u_n^{(h)}$. Как правило, мы будем так же поступать в аналогичных случаях и в дальнейшем.

В случае задачи (4) для отыскания сеточной функции $u^{(h)}$, приближенно совпадающей с искомой таблицей решения $[u]_h$, можно воспользоваться разностной схемой

$$\begin{gathered}\frac{u_{n+1}-2u_n+u_{n-1}}{h^2} - (1+x_n^2)\, u_n = \sqrt{x_n+1}, \\ u_0 = 2, \quad u_N = 1, \quad n = 1, 2, ..., N-1.\end{gathered} \tag{7}$$

Эта схема возникает в результате замены в точках сетки производной d^2u/dx^2, входящей в дифференциальное уравнение, по приближенной формуле

$$\frac{d^2u}{dx^2} \approx \frac{u(x+h)-2u(x)+u(x-h)}{h^2}. \tag{8}$$

Для вычисления решения $u^{(h)}$ этой задачи можно воспользоваться алгоритмом исключения (прогонки), описанным в § 4 гл. 4.

Выпишем еще разностную схему, пригодную для вычисления решения задачи (5):

$$\begin{gathered}\frac{v_{n+1}-v_n}{h} + x_n v_n w_n = x_n^3 - 3x_n + 1, \\ \frac{w_{n+1}-w_n}{h} + \frac{1}{1+x_n^2}\,(v_n+w_n) = \cos^2 x_n, \quad n = 0, 1, ..., N-1, \\ v_0 = 1, \quad w_0 = -3.\end{gathered} \tag{9}$$

Здесь $u_0^{(h)} = \begin{bmatrix} v_0 \\ w_0 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$ задано. При $n = 0$ из уравнений (9) можно найти $u_1^{(h)} = \begin{bmatrix} v_1 \\ w_1 \end{bmatrix}$. Вообще, зная $u_k = \begin{bmatrix} v_k \\ w_k \end{bmatrix}$ $(k = 0, 1, ..., n)$, можно при $k = n$ вычислить $u_{n+1} = \begin{bmatrix} v_{n+1} \\ w_{n+1} \end{bmatrix}$.

В рассмотренных примерах сетка D_h состоит из удаленных друг от друга на расстояние h точек. Ясно, что можно было бы расположить $N+1$ точек сетки D_h $(h = 1/N)$ на отрезке $[0, 1]$ не равномерно, а так, чтобы выполнялись условия

$$x_0 = 0, \quad x_1 = x_0 + h_0, \quad x_2 = x_1 + h_1, \ ..., \ x_{n+1} = x_n + h_n, \ ..., \ x_N = 1,$$

где $h_n > 0$ $(n = 0, 1, ..., N-1)$ — неравные между собой числа, однако такие, что $\max_n h_n \to 0$ при $N \to \infty$. Выбором расположения узлов сетки D_h можно добиться того, чтобы искомая таблица $[u]_h$ решения $u(x)$ была подробнее при данном фиксированном N (или $h = N^{-1}$) на тех участках, где $u(x)$ более быстро изменяется. Такие участки иногда бывают заранее известны из физики или из предварительных грубых расчетов. Информация о скорости изменения $u(x)$ выявляется также в ходе последовательного вычисления $u_1, u_2, ..., u_n$ и может быть учтена при выборе следующего узла сетки x_{n+1}.

Мы ограничимся приведенными примерами для иллюстрации понятия сетки и искомой сеточной функции (или вектор-функции) — таблицы значений решения $[u]_h$.

Мы ставим вопрос о вычислении сеточной функции $u^{(h)}$ потому, что при измельчении сетки, т. е. при $h \to 0$, она является все более подробной таблицей искомого решения $u(x)$ и дает о нем все более полное представление. Пользуясь интерполяцией, можно было бы с возрастающей при $h \to 0$ точностью восстановить решение u всюду в области D. Ясно, что точность, с которой это можно сделать при заданных фиксированном числе и расположении узлов сетки D_h, зависит от дополнительно известных сведений о решении типа оценок для его производных, а также от расположения узлов сетки.

Ограничимся этими беглыми замечаниями о восстановлении функции u по ее таблице $u^{(h)}$. Подробное рассмотрение вопросов восстановления функции по ее таблице составляет предмет теории интерполяции. Мы будем заниматься только задачей вычисления таблицы $u^{(h)}$. Поэтому условимся считать, что задача (1) решена точно, если найдена сеточная функция $[u]_h$. Однако нам не удастся вычислить ее точно. Вместо сеточной функции $[u]_h$ будем искать другую сеточную функцию $u^{(h)}$, которая сходится к $[u]_h$ при измельчении сетки. Для этой цели можно использовать разностные уравнения.

2. Сходящиеся разностные схемы. Способами построения и исследования сходящихся разностных схем мы будем заниматься на протяжении всей главы. Однако прежде надо придать точный смысл самому требованию сходимости $u^{(h)} \to [u]_h$, которое мы будем предъявлять к разностным схемам. Для этого рассмотрим линейное нормированное пространство функций, определенных на сетке D_h. Норма $\|u^{(h)}\|_{U_h}$ сеточной функции $u^{(h)} \in U_h$ есть неотрицательное число, которое принимается за меру отклонения функции $u^{(h)}$ от тождественного нуля.

Норма может быть определена различными способами. Можно, например, принять за норму функции точную верхнюю грань модуля ее значений в точках сетки, положив

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|. \qquad (10)$$

Если $u^{(h)} = \begin{bmatrix} v_k \\ w_k \end{bmatrix}$ $(k = 0, 1, ..., N)$, как в (9), то за норму, аналогичную (10), можно принять верхнюю грань модулей обеих функций v_k, w_k на соответствующей им сетке.

Если $u^{(h)}$ состоит из функций, определенных на сетке $x_n = nh$ $(n = 0, 1, ..., N)$, то часто используют норму, определенную равенством

$$\|u^{(h)}\|_{U_h} = \Big(h \sum_{n=0}^{N} u_n^2\Big)^{1/2}.$$

Эта норма аналогична норме

$$\|u(x)\| = \Big(\int_0^1 u^2 dx\Big)^{1/2}$$

для функций $u(x)$ с интегрируемым на отрезке $0 \leqslant x \leqslant 1$ квадратом.

Всюду, где не оговорено противное, мы будем пользоваться нормой (10).

После того как введено нормированное пространство U_h, приобретает смысл понятие отклонения одной функции от другой. Если $a^{(h)}$, $b^{(h)}$ — произвольные сеточные функции из U_h, то мерой их отклонения друг от друга считается норма их разности, т. е. число

$$\|b^{(h)} - a^{(h)}\|_{U_h}.$$

Теперь можно перейти к строгому определению сходящейся разностной схемы.

Пусть для приближенного вычисления решения дифференциальной краевой задачи (1), т. е. для приближенного вычисления сеточной функции $[u]_h$ на основе использования равенства (1), составлена некоторая система уравнений, которую будем символически записывать аналогично уравнению (1) в форме равенства

$$L_h u^{(h)} = f^{(h)}. \tag{11}$$

Примерами могут служить разностные схемы (6), (7), (9) для дифференциальных краевых задач (2), (4), (5) соответственно. Для записи схемы (6) в форме (11) можно положить

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} + \dfrac{nh}{1 + u_n^2}, & n = 0, 1, ..., N-1, \\ u_0, \end{cases}$$

$$f^{(h)} = \begin{cases} \cos nh, & n = 0, 1, ..., N-1, \\ 3. \end{cases}$$

Схема (7) запишется в форме (11), если принять

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + [1 - (nh)^2]u_n, & n = 1, 2, ..., N-1, \\ u_0, \\ u_N, \end{cases}$$

$$f^{(h)} = \begin{cases} \sqrt{1+nh}, & n = 1, 2, ..., N-1, \\ 2, \\ 1. \end{cases}$$

Запишем еще в виде (11) схему (9), приняв

$$L_h u^{(h)} = L_h \begin{bmatrix} v^{(h)} \\ w^{(h)} \end{bmatrix} =$$

$$= \begin{cases} \dfrac{v_{n+1} - v_n}{h} + (nh)v_n w_n, & n = 0, 1, ..., N-1, \\ \dfrac{w_{n+1} - w_n}{h} + \dfrac{1}{1+(nh)^2}(v_n + w_n), & n = 0, 1, ..., N-1, \\ v_0, \\ w_0, \end{cases}$$

$$f^{(h)} = \begin{cases} (nh)^2 - 3nh + 1, & n = 0, 1, ..., N-1, \\ \cos^2 nh, & n = 0, 1, ..., N-1, \\ 1, \\ -3. \end{cases}$$

Система (11), как видим, зависит от h и должна быть выписана для всех тех h, для которых рассматриваются сетка D_h и сеточная функция $[u]_h$. Таким образом, разностная краевая задача (11) — это не одна система, а семейство систем, зависящих от параметра h. Будем предполагать, что при каждом рассматриваемом достаточно малом h существует решение $u^{(h)}$ задачи (11), принадлежащее пространству U_h.

Будем говорить, что решение $u^{(h)}$ разностной краевой задачи (11) при измельчении сетки *сходится* к решению дифференциальной краевой задачи (1), если

$$\|[u]_h - u^{(h)}\|_{U_h} \to 0, \quad h \to 0. \tag{12}$$

Если сверх того выполнено неравенство

$$\|[u]_h - u^{(h)}\|_{U_h} \leqslant ch^k, \tag{13}$$

где $c > 0$, $k > 0$ — некоторые постоянные, не зависящие от h, то будем говорить, что имеет место *сходимость порядка* h^k или что *разностная схема имеет k-й порядок точности.*

Обладание свойством сходимости является фундаментальным требованием, которое предъявляется к разностной схеме (11) для численного решения дифференциальной краевой задачи (1). Если оно имеет место, то с помощью разностной схемы (11) можно вычислить решение $[u]_h$ с любой наперед заданной точностью, выбирая для этого h достаточно малым. Мы точно сформулировали понятие сходимости

и подошли к центральному вопросу о том, как построить сходящуюся разностную схему (11) для вычисления решения дифференциальной краевой задачи (1). Приведенные выше примеры схем дополняют рассмотренные в § 1 и дают представление о простейшем способе построения таких схем: следует выбрать сетку и заменить производные разностными отношениями. Однако для одной и той же дифференциальной краевой задачи можно получить различные разностные схемы (11), по-разному выбирая сетку D_h и по-разному заменяя производные приближающими их разностными отношениями.

3. Проверка сходимости разностной схемы. Не будем пока заниматься построением разностных схем и поставим задачу несколько иначе. Пусть разностная схема $L_h u^{(h)} = f^{(h)}$, позволяющая надеяться, что сходимость

$$\|[u]_h - u^{(h)}\|_{U_h} \to 0, \quad h \to 0,$$

имеет место, на основании тех или иных соображений уже построена. Как проверить, является ли она в самом деле сходящейся?

Предположим, что разностная задача (11) имеет единственное решение $u^{(h)}$. Если бы при подстановке в левую часть (11) вместо $u^{(h)}$ сеточной функции $[u]_h$ равенство (11) оказалось в точности выполненным, то в силу единственности решения имело бы место равенство $u^{(h)} = [u]_h$, идеальное с точки зрения сходимости. Это означало бы, что решение $u^{(h)}$ разностной задачи $L_h u^{(h)} = f^{(h)}$ совпадает с искомой сеточной функцией $[u]_h$, которую мы условились считать точным решением.

Однако, как правило, систему (11) не удается выбрать так, чтобы $[u]_h$ в точности ей удовлетворяла. При подстановке $[u]_h$ в уравнение (11) возникает некоторая невязка:

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}. \tag{14}$$

Если эта невязка $\delta f^{(h)}$ стремится к нулю при $h \to 0$, так что $[u]_h$ удовлетворяет уравнению (11) все точнее, то будем говорить, что разностная схема $L_h u^{(h)} = f^{(h)}$ *аппроксимирует* дифференциальную краевую задачу $Lu = f$ на решении u последней.

В случае аппроксимации можно считать, что уравнение (14), которому удовлетворяет $[u]_h$, получается из уравнения (11) путем прибавления некоторой малой (при малом h) добавки $\delta f^{(h)}$ к правой части $f^{(h)}$. Следовательно, если решение $u^{(h)}$ задачи (11) устойчиво относительно возмущения правой части $f^{(h)}$, т. е. мало изменяется при малом изменении правой части, то решение $u^{(h)}$ задачи (11) и решение $[u]_h$ задачи (14) различаются мало, так что из аппроксимации $\delta f^{(h)} \to 0$ следует сходимость

$$u^{(h)} \to [u]_h, \quad h \to 0.$$

Намеченный нами путь проверки сходимости (12) состоит в том, чтобы разбить этот трудный вопрос на два более простых: сначала проверить, имеет ли место аппроксимация задачи (1) задачей (11), а затем выяснить, устойчива ли задача (11). В этом содержится и указание на способы построения сходящихся разностных схем для численного решения задачи (1): надо строить аппроксимирующую ее разностную схему; из многих возможных способов аппроксимации надо выбирать такие, при которых разностные схемы оказываются устойчивыми.

Изложенный общий план исследования сходимости, естественно, предполагает, что введены математически строгие понятия аппроксимации и устойчивости, позволяющие доказать теорему о том, что из аппроксимации и устойчивости следует сходимость. Намеченные выше определения аппроксимации и устойчивости не являются строгими. Для определения аппроксимации надо еще уточнить, что такое невязка $\delta f^{(h)}$ в общем случае и что такое ее величина, а для определения устойчивости — придать точный смысл словам "малому возмущению правой части соответствует малое возмущение решения разностной задачи $L_h u^{(h)} = f^{(h)}$". Строгим определениям понятий аппроксимации и устойчивости мы посвятим отдельные параграфы.

§ 2. Аппроксимация дифференциальной краевой задачи разностной схемой

1. Невязка $\boldsymbol{\delta f^{(h)}}$. Придадим точный смысл понятию аппроксимации дифференциальной краевой задачи (1) из § 1

$$Lu = f \tag{1}$$

на решении u этой задачи разностной схемой (11) из § 1

$$L_h u^{(h)} = f^{(h)}. \tag{2}$$

Для этого надо уточнить, что такое невязка $\delta f^{(h)}$ в следующем равенстве:

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}, \tag{3}$$

возникающая при подстановке сеточной функции $[u]_h$ — таблицы искомого решения $u(x)$ — в уравнение (2), а также, что такое ее величина. Стремление величины невязки $\delta f^{(h)}$ к нулю при $h \to 0$ мы и примем затем за определение аппроксимации.

Начнем с рассмотрения примера разностной схемы для численного решения дифференциальной краевой задачи

$$\begin{gathered} \frac{d^2u}{dx^2} + a(x)\frac{du}{dx} + b(x)u = \cos x, \quad 0 \leqslant x \leqslant 1, \\ u(0) = 1, \quad u'(0) = 2. \end{gathered} \tag{4}$$

За сетку D_h по-прежнему примем совокупность точек $x_n = nh$ ($n = 0, 1, ..., N$). В качестве разностной схемы для приближенного вычисления $[u]_h$ воспользуемся совокупностью равенств

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + a(x_n)\frac{u_{n+1} - u_{n-1}}{2h} + b(x_n)u_n = \cos x_n,$$
$$n = 1, 2, ..., N-1, \tag{5}$$
$$u_0 = 1, \quad \frac{u_1 - u_0}{h} = 2,$$

возникающей при замене производных в (4) по приближенным формулам

$$\begin{aligned} \frac{d^2u(x)}{dx^2} &\approx \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}, \\ \frac{du(x)}{dx} &\approx \frac{u(x+h) - u(x-h)}{2h}, \\ \frac{du(0)}{dx} &\approx \frac{u(h) - u(0)}{h}. \end{aligned} \tag{6}$$

Разностная схема (5) записывается в форме (2), если обозначить

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - 2u_n + u_{n-1}}{h^2} + a(nh)\,\dfrac{u_{n+1} - u_{n-1}}{2h} + b(nh)\,u_n, \\ \qquad\qquad\qquad\qquad\qquad\qquad n = 1, 2, ..., N-1, \\ u_0, \\ \dfrac{u_1 - u_0}{h}, \end{cases}$$

$$f^{(h)} = \begin{cases} \cos nh, & n = 1, 2, ..., N-1, \\ 1, \\ 2. \end{cases}$$

Для вычисления и оценки величины невязки $\delta f^{(h)}$, возникающей при подстановке $[u]_h$ в уравнение (2), уточним формулы (6). По формуле Тейлора имеем

$$\begin{aligned} u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2!}\,u''(x) + \frac{h^3}{6}\,u'''(\xi_1), \\ u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2!}\,u''(x) - \frac{h^3}{6}\,u'''(\xi_2), \\ u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2!}\,u''(x) + \frac{h^3}{3!}\,u'''(x) + \frac{h^4}{24}\,u^{(4)}(\xi_3), \\ u(x-h) &= u(x) - hu'(x) + \frac{h^2}{2}\,u''(x) - \frac{h^3}{3!}\,u'''(x) + \frac{h^4}{24}\,u^{(4)}(\xi_4), \\ u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2!}\,u''(\xi_5). \end{aligned} \tag{7}$$

Здесь ξ_1, ξ_2, ξ_3, ξ_4, ξ_5 — некоторые промежуточные точки отрезка

$[x-h, x+h]$. Отсюда

$$\begin{aligned} \frac{u(x+h)-u(x-h)}{2h} &= u'(x) + \frac{h^2}{12}\,[u^{(3)}(\xi_1) + u^{(3)}(\xi_2)], \\ \frac{u(x+h)-2u(x)+u(x-h)}{h^2} &= u''(x) + \frac{h^2}{24}\,[u^{(4)}(\xi_3) + u^{(4)}(\xi_5)], \\ \frac{u(x+h)-u(x)}{h} &= u'(x) + \frac{h}{2}\,u''(\xi_5). \end{aligned} \tag{8}$$

2. Вычисление невязки. Будем считать, что решение $u(x)$ задачи (4) имеет ограниченные производные до четвертого порядка. В силу формул (8) можно написать

$$\frac{u(x+h)-2u(x)+u(x-h)}{h^2} + a(x)\,\frac{u(x+h)-u(x-h)}{2h} + b(x)\,u(x) =$$
$$= \frac{d^2u(x)}{dx^2} + a(x)\,\frac{du(x)}{dx} + b(x)\,u(x) +$$
$$+ h^2\Big[\frac{u^{(4)}(\xi_3) + u^4(\xi_4)}{24} + a(x)\,\frac{u^{(3)}(\xi_1) + u^3(\xi_2)}{12}\Big],$$
$$\xi_j \in [x-h, x+h].$$

Поэтому выражение

$$L_h[u]_h = \begin{cases} \dfrac{u(x_n+h)-2u(x_n)+u(x_n-h)}{h^2} + \\ \qquad + a(x_n)\,\dfrac{u(x_n+h)-u(x_n-h)}{2h} + b(x_n)\,u(x_n), \\ \qquad\qquad n = 1, 2, ..., N-1, \\ u(0), \\ \dfrac{u(h)-u(0)}{h} \end{cases}$$

можно переписать так:

$$L_h[u]_h = \begin{cases} \cos x_n + h^2\Big[\dfrac{u^{(4)}(\xi_3)+u^{(4)}(\xi_4)}{24} + a(x_n)\,\dfrac{u^{(3)}(\xi_1)+u^{(3)}(\xi_2)}{12}\Big], \\ \qquad\qquad n = 1, 2, ..., N-1, \\ 1 + 0, \\ 2 + h\dfrac{u''(\xi_5)}{2}, \end{cases}$$

или

$$L_h[u]_h = f^{(h)} + \delta f^{(h)},$$

где

$$\delta f^{(h)}\big|_n = \begin{cases} h^2\Big[\dfrac{u^{(4)}(\xi_3)+u^{(4)}(\xi_4)}{24} + a(x_n)\,\dfrac{u^{(3)}(\xi_1)+u^{(3)}(\xi_2)}{12}\Big], \\ \qquad\qquad n = 1, 2, ..., N-1, \\ 0, \\ h\dfrac{u''(\xi_5)}{2}. \end{cases} \tag{9}$$

Удобно считать, что $f^{(h)}$, $\delta^{(h)}$ принадлежат линейному нормированному пространству F_h, которое состоит из элементов вида

$$g^{(h)} = \begin{cases} \varphi_n, & n = 1, 2, \ldots, N-1, \\ \psi_0, \\ \psi_1, \end{cases} \tag{10}$$

где $\varphi_1, \varphi_2, \ldots, \varphi_{N-1}$, а также ψ_0, ψ_1 — произвольная упорядоченная система чисел; можно считать, что $g^{(h)}$ — это совокупность сеточной функции φ_n $(n = 1, 2, \ldots, N-1)$ и упорядоченной пары чисел ψ_0, ψ_1. Сложение двух элементов пространства F_h и умножение элементов $g^{(h)}$ на числа производятся покомпонентно. Ясно, что в рассматриваемом примере F_h есть $(N+1)$-мерное линейное пространство. Норма в F_h может быть введена многими способами. Если ввести в F_h норму равенством

$$\|g^{(h)}\|_{F_h} = \max\{|\psi_0|, |\psi_1|, \max_n |\varphi_n|\},$$

т. е. принять за норму максимум абсолютных величин всех компонент вектора $g^{(h)}$, то в силу (9) получим

$$\|\delta f^{(h)}\|_{F_h} \leqslant ch, \tag{11}$$

где c — некоторая постоянная, зависящая от $u(x)$, но не зависящая от h. Из этого неравенства следует стремление невязки $\delta f^{(h)}$ к нулю при $h \to 0$.

В уравнении $L_h u^{(h)} = f^{(h)}$, подробно записанном равенствами (5), которое мы рассмотрели в качестве примера, на L_h можно смотреть как на оператор. Этот оператор каждой сеточной функции $v^{(h)} = \{v_n\}$ $(n = 0, 1, \ldots, N)$ из линейного пространства U_h функций, определенных на сетке D_h, ставит в соответствие некоторый элемент $g^{(h)}$ вида (10) из линейного пространства F_h по формуле

$$L_h v^{(h)} = \begin{cases} \dfrac{v_{n+1} - 2v_n + v_{n-1}}{h^2} + a(x_n)\,\dfrac{v_{n+1} - v_{n-1}}{2h} + b(x_n)v_n, \\ v_0, \\ \dfrac{v_1 - v_0}{h}. \end{cases}$$

Условимся и в общем случае разностной краевой задачи (2) считать, что правые части тех скалярных уравнений, которые в совокупности записаны символическим равенством

$$L_h u^{(h)} = f^{(h)},$$

являются компонентами вектора $f^{(h)}$ из некоторого линейного нормированного пространства F_h. Тогда на L_h можно смотреть как на оператор, ставящий в соответствие каждой сеточной функции $u^{(h)}$

из U_h некоторый элемент $f^{(h)}$ из F_h. В таком случае имеет смысл выражение $L_h[u]_h$, возникающее в результате применения оператора L_h к сеточной функции $[u]_h$ из U_h и являющееся элементом пространства F_h. Невязка $\delta f^{(h)} = L_h[u]_h - f^{(h)}$ принадлежит пространству F_h как разность двух элементов этого пространства. Под величиной невязки следует понимать $\|\delta f^{(h)}\|_{F_h}$.

3. Аппроксимация порядка h^k.

Определение. Будем говорить, что разностная схема $L_h u^{(h)} = f^{(h)}$ *аппроксимирует задачу* $Lu = f$ *на решении* u, если $\|\delta f^{(h)}\|_{F_h} \to 0$ при $h \to 0$. Если сверх того имеет место неравенство

$$\|\delta f^{(h)}\|_{F_h} \leqslant ch^k,$$

где $c > 0$, $k > 0$ — некоторые постоянные, то будем говорить, что имеет место *аппроксимация порядка* h^k или *порядка* k *относительно величины* h.

То обстоятельство, что u является решением задачи (1), дает информацию о функции u, которую можно использовать для построения системы (2), а также для проверки факта аппроксимации. Поэтому в определении аппроксимации мы и упоминаем задачу (1). Однако подчеркнем, что приведенное определение аппроксимации задачи $Lu = f$ на решении u разностной схемой $L_h u^{(h)} = f^{(h)}$ само по себе не опирается на равенство $Lu = f$ для функции u. Можно было бы говорить просто о том, что схема $L_h u^{(h)} = f^{(h)}$ соответствует с порядком h^k функции u, не вникая в происхождение этой функции. В частности, если функция u является одновременно решением двух совсем различных задач вида (1), то одна и та же разностная схема $L_h u^{(h)} = f^{(h)}$ одновременно аппроксимирует или не аппроксимирует каждую из этих задач на их общем решении $u(x)$.

4. Примеры.

Пример 1. Разностная схема (5) в силу оценки (11) аппроксимирует задачу (4) с первым порядком относительно h. Разностную схему (5) легко усовершенствовать так, чтобы аппроксимация стала порядка h^2. Для этого заметим, что все компоненты вектора $\delta f^{(h)}$, кроме последней, стремятся к нулю, как h^2 (предпоследняя даже в точности равна нулю). Только последняя компонента вектора $\delta f^{(h)}$, т. е. невязка от подстановки $u(x)$ в последнее уравнение $(u_1 - u_0)/h = 2$ системы (5), стремится к нулю медленнее, а именно как первая степень h. Это досадное обстоятельство легко устранить.

По формуле Тейлора

$$\frac{u(h) - u(0)}{h} = u'(0) + \frac{h}{2}\,u''(0) + \frac{h^2}{3!}\,u'''(\xi) = 2 + \frac{h}{2}\,u''(0) + \frac{h^2}{6}\,u'''(\xi),$$
$$0 \leqslant \xi < h.$$

Но из дифференциального уравнения (4) находим

$$u''(0) = -a(0)u'(0) - b(0)u(0) + \cos 0 = -2a(0) - b(0) + 1.$$

Поэтому, заменив последнее равенство (5) равенством

$$\frac{u_1 - u_0}{h} = 2 - \frac{h}{2}\,[2a(0) + b(0) - 1], \tag{12}$$

получим для $f^{(h)}$ вместо (7) выражение

$$f^{(h)} = \begin{cases} \cos x_n, \\ 1, \\ 2 - \dfrac{h}{2}\,[2a(0) + b(0) - 1]. \end{cases}$$

Тогда окажется, что

$$\delta f^{(h)} = \begin{cases} \dfrac{h^2}{12}\Big[\dfrac{u^{(4)}(\xi_3) + u^{(4)}(\xi_4)}{2} + (u'''(\xi_1) + u'''(\xi_2))\Big], \\ 0, \\ \dfrac{h^2}{6}\,u'''(\xi), \end{cases}$$

$\|\delta f^{(h)}\|_{F_h} \leqslant ch^2$, где c — некоторая постоянная, не зависящая от h. Порядок аппроксимации станет вторым относительно h.

Подчеркнем, что для построения разностного граничного условия (12) мы использовали не только граничные условия задачи (4), но и само дифференциальное уравнение. Можно считать, что мы использовали граничное условие

$$u''(x) + a(x)u'(x) + b(x)u(x)|_{x=0} = \cos x|_{x=0},$$

которое является следствием дифференциального уравнения.

Пример 2. Выясним, каков порядок аппроксимации, которым обладает разностная схема

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_{n-1}}{2h} + Au_n = 1 + x_n^2, \quad n = 1, 2, ..., N-1, \\ u_0 = b, \\ u_1 = b \end{cases} \tag{13}$$

на решении u задачи

$$\frac{du}{dx} + Au = 1 + x^2, \quad u(0) = b. \tag{14}$$

Роль $f^{(h)}$ здесь играет

$$f^{(h)} = \begin{cases} 1 + x_n^2, \quad n = 1, 2, ..., N-1, \\ b, \\ b. \end{cases}$$

Далее,

$$L_h[u]_h = \begin{cases} \dfrac{u(x_{n+1}) - u(x_{n-1})}{2h} + Au(x_n), \quad n = 1, 2, ..., N-1, \\ u(0), \\ u(h), \end{cases}$$

или

$$L_h[u]_h = \begin{cases} \left[\dfrac{du(x_n)}{dx} + Au(x_n)\right] + \dfrac{h^2}{6}\, u'''(\xi_n), \quad n = 1, 2, ..., N-1, \\ u(0), \\ u(0) + h\, \dfrac{du(\xi_0)}{dx}, \quad 0 \leqslant \xi_0 \leqslant h. \end{cases}$$

Поскольку для решения $u(x)$ выполнено равенство

$$\frac{du(x_n)}{dx} + Au(x_n) = 1 + x_n^2,$$

то невязка $\delta f^{(h)}$ имеет вид

$$\delta f^{(h)} = \begin{cases} \dfrac{h^2}{6}\, u'''(\xi_n), \quad n = 1, 2, ..., N-1, \\ 0, \\ hu'(\xi_0). \end{cases}$$

Аппроксимация задачи (14) схемой (13) имеет первый относительно h порядок. Бросается в глаза, что компоненты невязки, как и в примере 1, имеют различные порядки относительно h. Разностное уравнение

$$\frac{u_{n+1} - u_{n-1}}{2h} + Au_n = 1 + x_n^2, \quad n = 1, 2, ..., N-1, \tag{15}$$

при подстановке $[u]_h$ удовлетворяется с невязкой $\dfrac{h^2}{6}\, u''(\xi_n)$ порядка h^2. Первое граничное условие $u_0 = b$ при подстановке $[u]_h$ выполнено точно, а второе условие $u_1 = b$ — с невязкой $hu'(\xi_0)$ порядка первой степени h.

5. Замена производных разностными отношениями. В рассмотренных примерах для получения разностных схем мы заменяли производные в дифференциальном уравнении разностными отношениями. Этот прием универсален и позволяет построить для любой дифференциальной краевой задачи, имеющей достаточно гладкое решение $u(x)$, разностную схему с любым наперед заданным порядком аппроксимации.

6. Другие способы построения разностных схем. Замена производных разностными отношениями не единственный, а часто и не лучший способ построения разностных схем. Некоторым другим способам, приводящим к наиболее употребительным разностным схемам, будет посвящен § 4. Здесь ограничимся примером.

Простейшая разностная схема

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - G(x_n, u_n) = 0, & n = 0, 1, ..., N-1, \\ u_0 = a, \quad Nh = 1, \end{cases}$$

называемая *схемой Эйлера*, аппроксимирует задачу

$$\frac{du}{dx} - G(x, y) = 0, \quad 0 \leqslant x \leqslant 1, \quad u(0) = a \tag{16}$$

с первым порядком относительно h. При известном u_n значение u_{n+1} вычисляется по формуле $u_{n+1} = u_n + hG(x_n, u_n)$. Схема

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - \dfrac{1}{2}\left[G(x_n, y_n) + G(x_{n+1}, \widetilde{u})\right] = 0, \\ u_0 = a, \end{cases}$$

где $\widetilde{u} = u_n + hG(x_n, u_n)$, называется *схемой Эйлера с пересчетом.* Она же является одной из схем Рунге–Кутта второго порядка аппроксимации, о которых будет подробно рассказано в § 4. Если u_n уже вычислено, то по схеме Эйлера вычисляем значение

$$\widetilde{u} = u_n + hG(x_n, u_n),$$

а потом осуществляем уточнение найденного $\widetilde{u}$, полагая

$$u_{n+1} = u_n + \frac{h}{2}\left[G(x_n, u_n) + G(x_{n+1}, \widetilde{u})\right].$$

На практике вычисление решения задачи Коши для обыкновенных дифференциальных уравнений без особенностей обычно производится по одной–двум довольно универсальным, хорошо апробированным схемам, для которых на современных компьютерах имеются стандартные программы.

Если приходится с очень большой точностью решать задачи специального вида, то применяются многочисленные специальные схемы, приспособленные именно для этих задач, но уступающие универсальным схемам при решении другого круга задач.

Задачи

1. Проверить, что схема Эйлера с пересчетом аппроксимирует задачу (16) на гладком решении $u(x)$ со вторым относительно h порядком.

2. Усовершенствовать второе начальное условие $u_1 = b$ в схеме (13) так, чтобы схема стала аппроксимировать с порядком h^2.

§ 3. Определение устойчивости разностной схемы. Сходимость как следствие аппроксимации и устойчивости

Выше мы построили ряд разностных схем, аппроксимирующих некоторые краевые задачи для обыкновенных дифференциальных уравнений. Можно показать, что эти схемы являются сходящимися, причем порядок сходимости совпадает с порядком аппроксимации.

Однако можно сконструировать примеры аппроксимирующих, но не сходящихся разностных схем. Легко проверить, что разностная схема

$$L_h u^{(h)} = \begin{cases} 4\,\dfrac{u_{n+1}-u_{n-1}}{2h} - 3\,\dfrac{u_{n+1}-u_n}{h} + Au_n = 0, \quad n = 1, 2, ..., N-1, \\ u_0 = b, \\ u_1 = be^{-Ah} \end{cases}$$

аппроксимирует задачу Коши

$$\frac{du}{dx} + Au = 0, \quad 0 \leqslant x \leqslant 1,$$
$$u(0) = b, \quad A = \text{const},$$

на решении u с первым порядком относительно h. Однако решение $u^{(h)}$, доставляемое этой разностной схемой, не стремится к $[u]_h$ и даже не остается ограниченным при $h \to 0$.

Действительно, общее решение этого разностного уравнения имеет вид

$$u_n = c_1 q_1^n + c_2 q_2^n,$$

где q_1, q_2 — корни квадратного уравнения $q^2 - (3 + Ah)q + 2 = 0$. Выбирая произвольные постоянные так, чтобы получить частное решение, удовлетворяющее заданным начальным условиям, получаем формулу

$$u_n = u_0\left[\frac{q_2}{q_2-q_1}\,q_1^n - \frac{q_1}{q_2-q_1}\,q_2^n\right] + u_1\left[-\frac{1}{q_2-q_1}\,q_1^n + \frac{1}{q_2-q_1}\,q_2^n\right].$$

Анализ этой формулы показывает, что $\max\limits_{0\leqslant nh\leqslant 1} |u_n| \to \infty$ при $h \to 0$.

Таким образом, аппроксимации, вообще говоря, недостаточно для сходимости. Нужна еще устойчивость.

1. Определение устойчивости. Пусть для приближенного вычисления решения и дифференциальной краевой задачи

$$Lu = f \tag{1}$$

составлена разностная схема

$$L_h u^{(h)} = f^{(h)}, \tag{2}$$

которая аппроксимирует задачу (1) на решении u с некоторым порядком h^k. Это значит, что невязка $\delta f^{(h)}$:

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}, \tag{3}$$

возникающая при подстановке таблицы $[u]_h$ решения u в уравнение (2), удовлетворяет оценке вида

$$\|\delta f^{(h)}\|_{F_h} \leqslant c_1 h^k, \tag{4}$$

где c — некоторая постоянная, не зависящая от h.

Определение 1. Будем называть разностную схему (2) *устойчивой*, если существуют числа $h_0 > 0, \delta > 0$ такие, что при любом $h < h_0$ и любом $\varepsilon^{(h)} \in F_h$, $\|\varepsilon^{(h)}\|_{F_h} < \delta$, разностная задача

$$L_h z^{(h)} = f^{(h)} + \varepsilon^{(h)}, \tag{5}$$

полученная из задачи (2) добавлением к правой части возмущения $\varepsilon^{(h)}$, имеет одно и только одно решение $z^{(h)}$, причем это решение отклоняется от решения $u^{(h)}$ невозмущенной задачи (2) на сеточную функцию $z^{(h)} - u^{(h)}$, удовлетворяющую оценке

$$\|z^{(h)} - u^{(h)}\|_{U_h} \leqslant c\,\|\varepsilon^{(h)}\|_{F_h}, \tag{6}$$

где c — некоторая постоянная, не зависящая от h.

В частности, неравенство (5) означает, что малое возмущение $\varepsilon^{(h)}$ правой части разностной схемы (2) вызывает равномерно относительно h малое возмущение $z^{(h)} - u^{(h)}$ решения.

Пусть оператор L_h, отображающий U_h в F_h, линейный. Тогда приведенное выше определение устойчивости равносильно следующему определению.

Определение 2. Будем называть разностную схему (2) с линейным оператором L_h *устойчивой*, если при любом $f^{(h)} \in F_h$ уравнение $L_h u^{(h)} = f^{(h)}$ имеет единственное решение $u^{(h)} \in U_h$, причем

$$\|u^{(h)}\|_{U_h} \leqslant c\,\|f^{(h)}\|_{F_h}, \tag{7}$$

где c — некоторая постоянная, не зависящая от h.

Докажем равносильность обоих определений устойчивости в случае линейного оператора.

Сначала установим, что из устойчивости разностной схемы (2) в смысле определения 2 следует устойчивость в смысле определения 1. Пусть линейная задача (2) при всех рассматриваемых $h_0 > 0$ и произвольном $f^{(h)} \in F_h$ имеет единственное решение, причем выполнена оценка (7). Вычитая из равенства (7) равенство (2), получаем

$$L_h(z^{(h)} - u^{(h)}) = \varepsilon^{(h)},$$

откуда в силу (7) следует оценка (6) при произвольном $\varepsilon^{(h)} \in F_h$, а значит, и устойчивость в смысле определения 1.

Покажем теперь, что устойчивость в смысле определения 1 влечет за собой устойчивость в смысле определения 2. В силу определения 1 при некоторых $h_0 > 0$, $\delta > 0$ и при произвольных $h < h_0$, $\varepsilon^{(h)} \in F_h$, $\|\varepsilon^{(h)}\|_{F_h} \leqslant \delta$, существуют и единственны решения уравнений

$$L_h z^{(h)} = f^{(h)} + \varepsilon^{(h)}, \quad L_h u^{(h)} = f^{(h)}.$$

Положим $w^{(h)} = z^{(h)} - u^{(h)}$ и вычтем эти равенства почленно. Получим

$$L_h w^{(h)} = \varepsilon^{(h)},$$

причем в силу (6)

$$\|w^{(h)}\|_{U_h} \leqslant c\,\|\varepsilon^{(h)}\|_{F_h}.$$

Очевидно, что, изменив обозначения решения и правой части уравнения $L_h w^{(h)} = \varepsilon^{(h)}$, последний результат можно сформулировать так: при произвольных $h < h_0$, $f^{(h)} \in F_h$, $\|f^{(h)}\|_{F_h} < \delta$, задача (2) имеет единственное решение $u^{(h)}$. Это решение удовлетворяет оценке (7). Однако в таком случае уравнение (2) имеет единственное решение $u^{(h)}$ и выполнена оценка (7) не только для всех $f^{(h)}$, удовлетворяющих оценке $\|f^{(h)}\|_{F_h} < \delta$, но и вообще для всех $f^{(h)} \in F_h$, т. е. имеет место устойчивость в смысле определения 2.

В самом деле, пусть $\|f^{(h)}\|_{F_h} > \delta$. Докажем однозначную разрешимость и оценку (7) в этом случае. Положим

$$u^{(h)} = \frac{2\|f^{(h)}\|_{F_h}}{\delta}\,\widetilde{u}^{(h)}, \quad f^{(h)} = \frac{2\|f^{(h)}\|_{F_h}}{\delta}\,\widetilde{f}^{(h)}.$$

Для $\widetilde{u}^{(h)}$ получим уравнение

$$L_h \widetilde{u}^{(h)} = f^{(h)},$$

причем

$$\|\widetilde{f}^{(h)}\|_{F_h} = \frac{\delta}{2\|f^{(h)}\|_{F_h}}\,\|f^{(h)}\|_{F_h} = \frac{\delta}{2} < \delta.$$

Поэтому уравнение $L_h \widetilde{u}^{(h)} = \overline{f}^{(h)}$ однозначно разрешимо, причем

$$\|\widetilde{u}^{(h)}\|_{U_h} \leqslant \|\widetilde{f}^{(h)}\|_{F_h}.$$

В силу формул, устанавливающих связь между $u^{(h)}$ и $\widetilde{u}^{(h)}$, а также между $f^{(h)}$ и $\widetilde{f}^{(h)}$, отсюда следует однозначная разрешимость задачи (2) и справедливость оценки (7) при произвольном рассматриваемом $f^{(h)} \in F_h$.

2. Зависимость между аппроксимацией, устойчивостью и сходимостью. Докажем теперь, что из аппроксимации и устойчивости следует сходимость.

Теорема 1. *Пусть разностная схема $L_h u^{(h)} = f^{(h)}$ аппроксимирует задачу $Lu = f$ на решении с порядком h^k и устойчива. Тогда решение $u^{(h)}$ разностной задачи $L_h u^{(h)} = f^{(h)}$ сходится к $[u]_h$, причем имеет место оценка*

$$\|[u]_h - u^{(h)}\|_{U_h} \leqslant cc_1 h^k, \tag{8}$$

где c, c_1 — числа, входящие в оценки (3), (5).

Доказательство. Положим $\varepsilon^{(h)} = \delta f^{(h)}$, $[u]_h = z^{(h)}$. Тогда оценка (5) примет вид

$$\|[u]_h - u^{(h)}\|_{U_h} \leqslant c\,\|\delta f^{(h)}\|_{F_h}.$$

Учитывая (3), сразу получаем доказываемое неравенство (7). □

В качестве иллюстрирующего примера докажем устойчивость разностной схемы Эйлера

$$\begin{gathered} \frac{u_{n+1} - u_n}{h} - G(x_n, u_n) = \varphi_n, \quad n = 0, 1, ..., N-1, \\ u_0 = \psi, \end{gathered} \tag{9}$$

$x_n = nh$, $h = 1/N$, для численного решения дифференциальной краевой задачи

$$\begin{gathered} \frac{du}{dx} - G(x, y)u = \varphi(x), \quad 0 \leqslant x \leqslant 1, \\ u|_{x=0} = \psi. \end{gathered} \tag{10}$$

Будем предполагать функцию $G(x, u)$ двух аргументов и функцию $\varphi(x)$ такими, что существует решение $u(x)$, имеющее ограниченную вторую производную. Кроме того, будем считать, что $G(x, u)$ имеет ограниченную производную по u:

$$|\partial G/\partial u| < M. \tag{11}$$

Положим $\|f^{(h)}\|_{F_h} = \max\{|\psi|, \max\limits_n |\varphi(x_n)|\}$.

Читателю рекомендуется проверить, что разностная схема (9) аппроксимирует задачу (10) на решении $u(x)$ с первым относительно h порядком. (Разностное уравнение соответствует задаче с первым порядком, а граничное условие $u_0 = \psi$ точно.)

Определим норму

$$\|u^{(h)}\|_{U_h} = \max_n |u_n|$$

и займемся проверкой устойчивости разностной схемы (9). Запишем ее в форме (2), положив

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - G(x_n, u_n), & n = 0, 1, ..., N-1, \\ u_0, \end{cases}$$

$$f^{(h)} = \begin{cases} \varphi(x_n), & n = 0, 1, ..., N-1, \\ \psi. \end{cases}$$

Задача

$$L_h z^{(h)} = f^{(h)} + \delta f^{(h)}$$

в подробной записи имеет вид

$$\begin{gathered}\frac{z_{n+1} - z_n}{h} - G(x_n, z_n) = \varphi(x_n) + \varepsilon_n, \quad n = 0, 1, ..., N-1, \\ z_0 = \psi + \varepsilon,\end{gathered} \tag{12}$$

где

$$\varepsilon^{(h)} = \begin{cases} \varepsilon_n, & n = 0, 1, ..., N-1, \\ \varepsilon. \end{cases}$$

Вычтем из уравнений (12) соответствующие уравнения (9) почленно. Обозначим $z_n - u_n = w_n$ и учтем, что

$$G(x_n, z_n) - G(x_n, u_n) = \frac{\partial G(x_n, \xi_n)}{\partial u}\, w_n \equiv M_n^{(h)} w_n,$$

где ξ_n — некоторое число, заключенное между числами z_n и w_n. Получим следующую систему уравнений для определения $w^{(h)} = \{w_0, w_1, ..., w_N\}$:

$$\begin{gathered}\frac{w_{n+1} - w_n}{h} - M_n^{(h)} w_n = \varepsilon_n, \quad n = 0, 1, ..., N-1, \\ w_0 = \varepsilon.\end{gathered} \tag{12}$$

Учитывая, что $|M_n^{(h)}| < M$ в силу (11) и что $nh \leqslant 1$, получаем

$$\begin{aligned}|w_{n+1}| = |(1 + hM_n^{(h)})\, w_n + h\varepsilon_n| \leqslant (1 + Mh)|w_n| + h\,|\varepsilon_n| \leqslant \\ \leqslant (1 + Mh)^2\, |w_{n-1}| + h(1 + Mh)|\varepsilon_{n-1}| + h\,|\varepsilon_n| \leqslant \\ \leqslant (1 + Mh)^2\, |w_{n-1}| + 2h(1 + Mh)\, \|\varepsilon^{(h)}\|_{F_h} \leqslant \\ \leqslant (1 + Mh)^3\, |w_{n-2}| + 3h(1 + Mh)^2\, \|\varepsilon^{(h)}\|_{F_h} \leqslant \\ \leqslant (1 + Mh)^{n+1}\, |w_0| + (n+1)\, h(1 + Mh)^n\, \|\varepsilon^{(h)}\|_{F_h} \leqslant \\ \leqslant 2(1 + Mh)^N \|\varepsilon^{(h)}\|_{F_h} \leqslant 2e^M \|\varepsilon^{(h)}\|_{F_h}.\end{aligned}$$

Из доказанного неравенства

$$|w_{n+1}| \leqslant 2e^M \|\varepsilon^{(h)}\|_{F_h}$$

следует оценка вида (6):

$$\|w^{(h)}\|_{U_h} \leqslant 2e^M \|\varepsilon^{(h)}\|_{F_h},$$

означающая устойчивость с постоянной $c = 2e^M$. В силу теоремы 1 разностная схема (9) является сходящейся с первым относительно h порядком.

Исследуем теперь сходимость разностной схемы (7) из § 1:

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} - (1 + x_n^2)\, u_n = \sqrt{1 + x_n}, \quad n = 1, 2, ..., N - 1, \quad (14)$$
$$u_0 = 2, \quad u_N = 1$$

для дифференциальной краевой задачи (4) из § 1. Аппроксимация со вторым относительно h порядком задачи (4) из § 1 задачей (14) благодаря формуле

$$\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = u''(x) + \frac{h^2}{12}\, u^{(4)}(\xi)$$

здесь очевидна.

Займемся проверкой устойчивости. Рассматриваемая задача линейна. Поэтому проверка устойчивости состоит в том, чтобы установить существование единственного решения задачи

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{h^2} - (1 + x_n^2)\, u_n = g_n, \quad n = 1, 2, ..., N - 1, \quad (15)$$
$$u_0 = \alpha, \quad u_N = \beta$$

при любых $\{g_n\}$, α, β и в том, чтобы получить оценку

$$\max |u_n| \leqslant c \max \big\{|\alpha|, |\beta|, \max |g_n|\big\}. \quad (16)$$

Задачу вида (15) мы рассматривали в связи с прогонкой. Для задачи вида

$$a_n u_{n-1} + b_n u_n + c_n u_{n+1} = g_n,$$
$$u_0 = \alpha, \quad u_N = \beta$$

в предположении

$$|b_n| > |a_n| + |c_n| + \delta, \quad \delta > 0,$$

были доказаны ее однозначная разрешимость и оценка

$$|u_n| \leqslant \max \Big\{|\alpha|, |\beta|, \frac{1}{\delta} \max |g_m|\Big\}. \quad (17)$$

В случае задачи (15)

$$a_n = 1/h^2, \quad c_n = 1/h^2, \quad b_n = 2/h^2 + 1 + x_n^2 \geqslant |a_n| + |b_n| + 1.$$

Поэтому оценка (17) влечет за собой оценку (16) с постоянной $c = 1$. Устойчивость доказана.

В заключение параграфа подчеркнем, что схема доказательства сходимости решения задачи $L_n u^{(h)} = f^{(h)}$ к решению задачи $Lu = f$ путем проверки аппроксимации и устойчивости носит общий характер. Под $Lu = f$ можно понимать любое функциональное уравнение, а не только краевую задачу для обыкновенного дифференциального

уравнения. Само по себе неважно, решением какой задачи является функция u. Уравнение $Lu = f$ используется только для конструирования разностного уравнения $L_h u^{(h)} = f^{(h)}$. Поясним эту мысль.

3. Сходящаяся разностная схема для интегрального уравнения. Построим и исследуем разностную схему для вычисления решения интегрального уравнения

$$Lu \equiv u(x) - \int_0^1 K(x,y)u(y)\,dy = f(x), \quad 0 \leqslant x \leqslant 1.$$

Будем предполагать, что $|K(x,y)| < \rho < 1$.

Зададим N, положим $h = 1/N$ и будем искать таблицу $[u]_h$ значений решения на сетке $x_n = nh$ $(n = 0, 1, ..., N)$. Для получения разностной схемы в равенстве

$$u(x_n) - \int_0^1 K(x_n, y)u(y)\,dy = f(x_n), \quad n = 0, 1, ..., N,$$

приближенно заменим интеграл суммой, пользуясь квадратурной формулой трапеций.

Напомним эту формулу. Для произвольной, дважды дифференцируемой на отрезке $0 \leqslant y \leqslant 1$ функции $\varphi(y)$ справедливо приближенное равенство

$$\int_0^1 \varphi(y)\,dy \approx h\Big(\frac{\varphi_0}{2} + \varphi_1 + ... + \varphi_{N-1} + \frac{\varphi_N}{2}\Big), \quad h = \frac{1}{N},$$

причем погрешность есть величина $O(h^2)$. После указанной замены интеграла получим

$$u_h - h\Big(\frac{K(x_n, 0)}{2}\,u_0 + K(x_n, h)\,u_1 + ... + K(x_n, y_{N-1})\,u_{N-1} + \\ + \frac{K(x_n, 1)}{2}\,u_N\Big) = f_n, \quad n = 0, 1, ..., N. \quad (18)$$

Выписанная система равенств записывается в форме $L_h u^{(h)} = f^{(h)}$, если положить

$$L_h u^{(h)} = \begin{cases} g_0, \\ g_1, \\ \dots \\ g_N, \end{cases} \qquad f^{(h)} = \begin{cases} f(0), \\ f(h), \\ \dots \\ f(Nh), \end{cases}$$

где

$$g_n = u_n - h\Big[\frac{K(x_n, 0)}{2}\,u_0 + K(x_n, h)\,u_1 + ... + \frac{K(x_n, 1)}{2}\,u_N\Big], \\ n = 0, 1, ..., N.$$

Построенная разностная схема $L_h u^{(h)} = f^{(h)}$ аппроксимирует задачу $Lu = f$ на решении u со вторым порядком относительно шага h, поскольку квадратурная формула трапеций имеет второй порядок точности.

Проверим устойчивость. Пусть $u^{(h)} = (u_0, u_1, ..., u_N)$ — какое-нибудь решение системы (18) и пусть u_s — одна из тех компонент решения, которые по модулю не меньше каждой из остальных:

$$|u_s| \geqslant |u_n|, \quad n = 0, 1, ..., N.$$

Из уравнения с номером $n = s$ системы (18) следует неравенство

$$|f(x_s)| = \left| u_s - h\Big(\frac{K(x_s,0)}{2}\, u_0 + K(x_s,h)\, u_1 + ... + \frac{K(x_s,1)}{2}\, u_N\Big) \right| \geqslant$$
$$\geqslant |u_s| - h\Big(\frac{\rho}{2} + \rho + ... + \rho + \frac{\rho}{2}\Big)|u_s| = (1 - nh\rho)|u_s| = (1-\rho)|u_s|.$$

Поэтому

$$\|u^{(h)}\|_{U_h} = \max_n |u_n| = |u_s| \leqslant \frac{1}{1-\rho}\,|f(x_s)| \leqslant \frac{1}{1-\rho}\,\|f^{(h)}\|_{F_h}. \tag{19}$$

В частности, при $f(x_n) \equiv 0$ отсюда следует, что система (17) не имеет нетривиальных решений, а следовательно, однозначно разрешима при любой правой части. Неравенство (18) означает устойчивость (6) с постоянной $c = 1/(1-\rho)$. Решение $u^{(h)}$ задачи $L_h u^{(h)} = f^{(h)}$ в силу теоремы о сходимости удовлетворяет неравенству

$$\|[u]_h - u^{(h)}\|_{U_h} = \max_n |u(mh) - u_n| \leqslant ch^2,$$

где c — некоторая постоянная.

§ 4. Схемы Рунге–Кутта

Изложим здесь некоторые употребительные разностные схемы решения задачи Коши для дифференциального уравнения первого порядка

$$\frac{du}{dx} - G(x,u) = 0, \quad 0 \leqslant x \leqslant 1, \quad u(0) = a. \tag{1}$$

В конце параграфа эти схемы будут перенесены на системы дифференциальных уравнений первого порядка, к которым сводится общий случай уравнений и систем любого порядка.

Выберем на отрезке $0 \leqslant x \leqslant 1$ сетку точек

$$0 = x_0 < x_1 < ... < x_{N-1} < x_N = 1, \quad x_n = nh, \quad h = 1/N,$$

и будем составлять разностные схемы для приближенного отыскания таблицы $[u]_h$ значений решения $u(x)$ на выбранной сетке.

С простейшей употребительной схемой мы уже встречались. Это — схема Эйлера

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - G(x_n, u_n) = 0, \quad n = 0, 1, ..., N-1, \\ u_0 = a, \end{cases} \tag{2}$$

обладающая первым порядком аппроксимации (и точности).

Вычисления по этой схеме имеют простой геометрический смысл. Если u_n уже вычислено, то вычисление

$$u_{n+1} = u_n + hG(x_n, u_n)$$

равносильно сдвигу из точки (x_n, u_n) в точку (x_{n+1}, u_{n+1}) на плоскости Oxu по касательной к интегральной кривой $u = u(x)$ дифференциального уравнения $du/dx - G(x, u) = 0$, проходящей через точку (x_n, u_n).

Среди схем более высокого порядка аппроксимации наиболее употребительны различные варианты схем Рунге–Кутта.

1. Схемы Рунге–Кутта. Пусть значение u_n приближенного решения в точке x_n уже найдено и требуется вычислить u_{n+1} в точке $x_{n+1} = x_n + h$. Задаем целое l и выписываем выражения

$$\begin{aligned} k_1 &= G(x_n, u_n), \\ k_2 &= G(x_n + \alpha h, u_n + \alpha h k_1), \\ k_3 &= G(x_n + \beta h, u_n + \beta h k_2), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ k_l &= G(x_n + \gamma h, u_n + \gamma h k_{l-1}). \end{aligned}$$

Затем полагаем

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - (p_1 k_1 + p_2 k_2 + ... + p_l k_l) = 0, \\ \qquad\qquad\qquad\qquad n = 0, 1, ..., N-1, \\ u_0 = a. \end{cases}$$

Коэффициенты $\alpha, \beta, ..., \gamma$, $p_1, p_2, ..., p_l$ подбираем так, чтобы получить при заданном l аппроксимацию возможно более высокого порядка. Зная u_n, можно вычислить $k_1, k_2, ..., k_l$, а затем $u_{n+1} = u_n + h(p_1 k_1 + p_2 k_2 + ... + p_l k_l)$.

Простейшей схемой Рунге–Кутта является схема Эйлера ($l = 1$).

Схема Рунге–Кутта

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - \dfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 0, \\ \qquad\qquad\qquad\qquad n = 0, 1, ..., N-1, \\ u_0 = a, \end{cases} \tag{3}$$

где

$$k_1 = G(x_n, u_n), \quad k_2 = G\Big(x_n + \frac{h}{2}, u_n + \frac{k_1 h}{2}\Big),$$
$$k_3 = G\Big(x_n + \frac{h}{2}, u_n + \frac{k_2 h}{2}\Big), \quad k_4 = G(x_n + h, u_n + k_3 h),$$

имеет четвертый порядок аппроксимации.

Схема Рунге–Кутта

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{n+1} - u_n}{h} - \Big[\dfrac{2\alpha - 1}{2\alpha}\, k_1 + \dfrac{1}{2\alpha}\, k_2\Big] = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad n = 0, 1, ..., N-1, \\ u_0 = a, \end{cases} \tag{4}$$

где $k_1 = G(x_n, u_n)$, $k_2 = G(x_n + \alpha h, u_n + \alpha h k_1)$, при любом фиксированном α имеет второй порядок аппроксимации.

Мы докажем только утверждение о схеме (4). Доказательство утверждения о схеме (3) аналогично, но более громоздко.

Решение $u(x)$ уравнения $u' = G(x, u)$ удовлетворяет тождествам

$$\frac{du}{dx} \equiv G(x, u(x)),$$
$$\frac{d^2 u}{dx^2} \equiv \frac{d}{dx}\, G(x, u) = \frac{\partial G}{\partial x} + \frac{\partial G}{\partial u}\, G.$$

Поэтому из формулы Тейлора

$$\frac{u(x_n + h) - u(x_n)}{h} = u'(x_n) + \frac{h}{2}\, u''(x_n) + O(h^2)$$

для решения $u(x)$ следует равенство

$$\frac{u(x_{n+1}) - u(x_n)}{h} - \Big[G + \frac{h}{2}\, \frac{\partial G}{\partial x} + \frac{\partial G}{\partial u}\, G\Big]_{\substack{x = x_n \\ u = u(x_n)}} = O(h^2). \tag{5}$$

Но, разлагая по h функцию двух переменных по формуле Тейлора и удерживая члены первой степени, получаем

$$\frac{2\alpha - 1}{2\alpha}\, k_1 + \frac{1}{2\alpha}\, k_2 =$$
$$= \frac{2\alpha - 1}{2\alpha}\, G + \frac{1}{2\alpha}\Big[G + \frac{\partial G}{\partial x}\, \alpha h + \frac{\partial G}{\partial u}\, \alpha h G + O(h^2)\Big]_{\substack{x = x_n \\ u = u(x_n)}} =$$
$$= G + \frac{h}{2}\Big(\frac{\partial G}{\partial x} + \frac{\partial G}{\partial u}\, G\Big)\Big|_{\substack{x = x_n \\ u = u(x_n)}} + O(h^2). \tag{6}$$

Поэтому при подстановке в левую часть равенства (4) вместо u_n, u_{n+1} соответственно значений $u(x_n)$, $u(x_{n+1})$ решения $u(x)$ получится выражение, совпадающее с левой частью равенства (5) с точностью

$O(h^2)$. Следовательно, это выражение имеет второй порядок относительно h. Поскольку значение $u_0 = a$ задано точно, этим завершается доказательство того, что схема (4) имеет второй порядок аппроксимации.

2. Обобщение на системы уравнений. Все описанные схемы численного решения задачи Коши для дифференциального уравнения первого порядка (1) автоматически переносятся на системы уравнений первого порядка. Для этого в записи (1) надо понимать под $u(x) = \mathbf{u}(x)$ и $G(x, u) = \mathbf{G}(x, \mathbf{u})$ вектор-функции, а под $a = \mathbf{a}$ — заданный вектор. Тогда схемы Рунге–Кутта (3), (4) сохранят смысл и останутся применимыми.

Например, система уравнений

$$\frac{dv}{dx} - (x + v^2 + \sin w) = 0,$$
$$\frac{dw}{dx} + xvw = 0,$$
$$v(0) = a_1, \quad w(0) = a_2$$

запишется в форме

$$\frac{d\mathbf{u}}{dx} - \mathbf{G}(x, \mathbf{u}) = 0,$$
$$\mathbf{u}(0) = \mathbf{a},$$

если положить

$$\mathbf{u}(x) = \begin{bmatrix} v(x) \\ w(x) \end{bmatrix}, \quad \mathbf{G}(x, \mathbf{u}) = \begin{bmatrix} x + v^2 + \sin w \\ -xvw \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}.$$

Формула для $\mathbf{u}_{n+1}$ в схеме Эйлера:

$$\mathbf{u}_{n+1} = \mathbf{u}_n + h\mathbf{G}(x_n, \mathbf{u}_n),$$

подробно запишется так:

$$v_{n+1} = v_n + h(x_n + v_n^2 + \sin w_n),$$
$$w_{n+1} = w_n + h(-x_n v_n w_n).$$

Все рассуждения о порядке аппроксимации в случае одного уравнения сохраняются и в случае системы уравнений. При этом в формуле (6) под производной вектора $\mathbf{G} = (G_1, G_2, ..., G_k)$ по вектору $\mathbf{u} = (u_1, u_2, ..., u_k)$ надо понимать матрицу

$$\partial\mathbf{G}/\partial\mathbf{u} = \begin{bmatrix} \partial G_1/\partial u_1 & \ldots & \partial G_1/\partial u_k \\ \ldots & \ldots & \ldots \\ \partial G_k/\partial u_1 & \ldots & \partial G_k/\partial u_k \end{bmatrix}.$$

Произвольная система дифференциальных уравнений, разрешенных относительно старших производных, сводится к системе уравне-

ний первого порядка $d\mathbf{u}/dx = \mathbf{G}(x, \mathbf{u})$ путем замены искомых функций. Как это делается, ясно из следующего примера. Задача

$$\frac{d^2v}{dx^2} + \sin(xv' + v^2 + w) = 0,$$
$$\frac{dw}{dx} + \sqrt{x^2 + v^2 + (v')^2 + w^2} = 0,$$
$$v(0) = a, \quad v'(0) = b, \quad w(0) = c$$

приводится к требуемому виду, если положить

$$u_1 = v(x), \quad u_2 = dv/dx, \quad u_3 = w(x).$$

Получим

$$\frac{du_1}{dx} - u_2 = 0,$$
$$\frac{du_2}{dx} + \sin(xu_2 + u_1^2 + u_3) = 0,$$
$$\frac{du_3}{dx} + \sqrt{x^2 + u_1^2 + u_2^2 + u_3^2} = 0,$$
$$u_1(0) = a, \quad u_2(0) = b, \quad u_3(0) = c.$$

Замечание. Разработаны разностные схемы типа схем Рунге–Кутта, применимые непосредственно для уравнения второго порядка и не требующие предварительного сведения этих уравнений к системам первого порядка.

Задача

Дифференциальное уравнение

$$\frac{dy}{dx} = \frac{16x - 5y + xy}{x + y - xy}$$

имеет особую точку (седло) при $x = y = 0$.

Численно построить сепаратрисы, объединяя аналитические средства для выхода из особой точки с расчетами с помощью разностных схем.

§ 5. Методы решения краевых задач

Примером краевой задачи является задача

$$y'' = f(x, y, y'), \quad 0 \leqslant x \leqslant 1, \qquad y(0) = Y_0, \quad y(1) = Y_1 \tag{1}$$

с граничными условиями на концах отрезка $0 \leqslant x \leqslant 1$, на котором надо найти решение $y(x)$. На этом примере схематически изложим некоторые способы численного решения краевых задач.

1. Метод стрельбы. В § 4 указаны удобные способы численного решения задачи Коши, т. е. задачи вида

$$y'' = f(x, y, y'), \quad 0 \leqslant x < 1,$$
$$y(0) = Y_0, \quad \left.\frac{dy}{dx}\right|_{x=0} = \operatorname{tg}\alpha, \tag{2}$$

где Y_0 — ордината точки $(0, Y_0)$, из которой выходит интегральная кривая, а α — угол наклона интегральной кривой к оси Ox при выходе из точки $(0, Y_0)$ (рис. 15). При фиксированном Y_0 решение задачи (2) имеет вид $y = y(x, \alpha)$. При $x = 1$ решение $y(x, \alpha)$ зависит только от α:

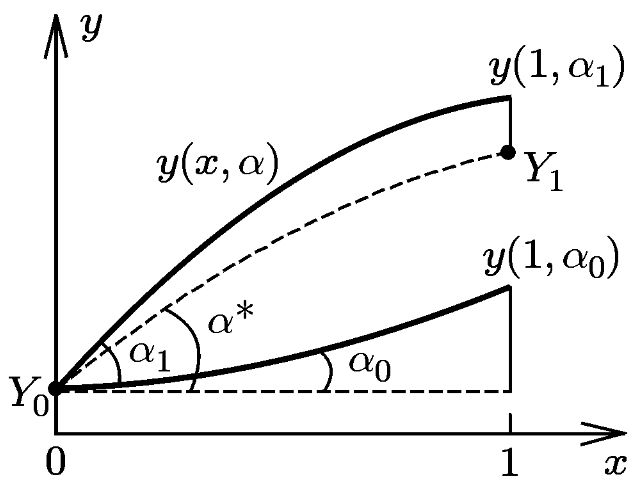


Рис. 15

$$y(x, \alpha)|_{x=1} = y(1, \alpha).$$

Используя указанное замечание о решении задачи Коши (2), мы можем переформулировать задачу (1) следующим образом: найти такой угол $\alpha = \alpha^*$, при котором интегральная кривая, выходящая из точки $(0, Y_0)$ под углом α к оси абсцисс, попадет в точку $(1, Y_1)$:

$$y(1, \alpha) = Y_1. \tag{3}$$

Решение задачи (2) при этом $\alpha = \alpha^*$ совпадает с искомым решением задачи (1). Дело сводится, таким образом, к решению уравнения (3). Уравнение (3) есть уравнение вида $F(\alpha) = 0$, где $F(\alpha) = y(1, \alpha) - Y_1$. Оно отличается от привычных уравнений лишь тем, что функция $F(\alpha)$ задана не аналитическим выражением, а с помощью алгоритма решения задачи (2). Сведение решения краевой задачи (1) к решению задачи Коши (2) и составляет сущность метода стрельбы.

Для решения уравнения (3) можно использовать метод деления отрезка пополам, метод хорд, метод касательных (метод Ньютона) и т. д. Например, при использовании метода деления отрезка пополам мы задаем α_0, α_1 так, чтобы разности $y(1, \alpha_0) - Y_1$, $y(1, \alpha_1) - Y_1$ имели разные знаки. Затем полагаем

$$\alpha_2 = (\alpha_0 + \alpha_1)/2.$$

Вычисляем $y(1, \alpha_2)$. Вычисляем затем α_3 по одной из формул

$$\alpha_3 = (\alpha_1 + \alpha_2)/2, \quad \alpha_3 = (\alpha_0 + \alpha_2)/2$$

в зависимости от того, имеют разности $y(1, \alpha_2) - Y_1$, $y(1, \alpha_1) - Y_1$ соответственно разные или одинаковые знаки. Затем вычисляем $y(1, \alpha_3)$. Процесс продолжается до тех пор, пока не будет достигнута требуемая точность $|y(1, \alpha_n) - Y_1| < \varepsilon$.

В случае использования метода хорд задаем α_0, α_1, а затем последующие α_k вычисляем по рекуррентной формуле

$$\alpha_{n+1} = \alpha_n - \frac{F(\alpha_n)}{F(\alpha_n) - F(\alpha_{n-1})}(\alpha_n - \alpha_{n-1}), \quad n = 1, 2, \ldots$$

Метод стрельбы, сводящий решение краевой задачи (1) к вычислению решений задачи Коши (2), хорошо работает в том случае, если решение $y(x, \alpha)$ "не слишком сильно" зависит от α. В противном случае он становится вычислительно неустойчивым, даже если решение задачи (1) зависит от входных данных "умеренно".

Поясним взятые в кавычки слова на примере следующей линейной краевой задачи:

$$\begin{gathered} y'' - a^2 y = 0, \quad 0 \leqslant x \leqslant 1, \\ y(0) = Y_0, \quad y(1) = Y_1 \end{gathered} \tag{1$'$}$$

при постоянном a^2. Выпишем решение этой задачи:

$$y(x) = \frac{e^{-ax} - e^{-a(2-x)}}{1 - e^{-2a}}\, Y_0 + \frac{e^{-\alpha(1-x)} - e^{-a(1+x)}}{1 - e^{-2a}}\, Y_1.$$

Коэффициенты при Y_0, Y_1 с ростом a остаются ограниченными на отрезке $0 \leqslant x \leqslant 1$ функциями; при всех $a > 0$ они не превосходят единицу. Поэтому небольшие погрешности при задании Y_0, Y_1 ведут к столь же небольшим погрешностям в решении.

Рассмотрим теперь задачу Коши

$$\begin{gathered} y'' - a^2 y = 0, \quad 0 \leqslant x \leqslant 1. \\ y(0) = Y_0, \quad y'(0) = \operatorname{tg}\alpha. \end{gathered} \tag{2$'$}$$

Ее решение имеет вид

$$y(x) = \frac{aY_0 + \operatorname{tg}\alpha}{2a}\, e^{ax} + \frac{aY_0 - \operatorname{tg}\alpha}{2a}\, e^{-ax}.$$

Если при задании $\operatorname{tg}\alpha$ допущена погрешность ε, то значение решения при $x = 1$ получит приращение

$$\Delta y(1) = \frac{\varepsilon}{2a}\, e^{a} - \frac{\varepsilon}{2a}\, e^{-a}. \tag{4}$$

При больших a вычитаемое в этом равенстве пренебрежимо мало, но коэффициент при ε в первом слагаемом становится большим. Поэтому метод стрельбы при решении задачи (1$'$), будучи формально приемлемой процедурой, при больших a становится практически непригодным.

2. Метод прогонки. Для решения краевой задачи

$$\begin{gathered} y'' - p(x)y = f(x), \quad 0 \leqslant x \leqslant 1, \\ y(0) = Y_0, \quad y(1) = Y_1 \end{gathered}$$

при $p(x) \gg 1$ можно воспользоваться разностной схемой

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p(x_n)y_n = f_n, \quad n = 1, 2, ..., N-1,$$
$$y(0) = Y_0, \quad y(1) = Y_1$$

и решать разностную задачу прогонкой. Условия применимости прогонки при $p(x) > 0$, как легко проверит читатель, выполнены.

3. Метод Ньютона. Метод стрельбы при решении хорошо поставленной краевой задачи может оказаться, как мы видели, неприменимым из-за вычислительной неустойчивости. Но метод прогонки даже формально можно применять только для решения линейных задач.

Метод Ньютона сводит решение нелинейной задачи к серии линейных задач и состоит в следующем. Пусть известна некоторая функция $y_0(x)$, удовлетворяющая граничным условиям (1) и грубо приближенно равная искомому решению $y(x)$. Положим

$$y(x) = y_0(x) + v(x), \tag{5}$$

где $v(x)$ — поправка к нулевому приближению $y_0(x)$. Подставим (5) в уравнение (1) и линеаризуем задачу, используя равенства

$$y''(x) = y_0''(x) + v''(x),$$
$$f(x, y_0 + v, y_0' + v_0') =$$
$$= f(x, y_0, y_0') + \frac{\partial f(x, y_0, y_0')}{\partial y} v + \frac{\partial f(x, y_0, y_0')}{\partial y'} v' + O(v^2 + |v'|^2).$$

Отбрасывая остаточный член $O(v^2 + |v'|^2)$, получаем линейную задачу для поправки $\overline{v}$

$$\overline{v}'' = p\overline{v}' + q\overline{v} + \varphi(x), \quad \overline{v}(0) = \overline{v}(1) = 0, \tag{6}$$

где

$$p = p(x) = \frac{\partial f(x, y_0, y_0')}{\partial y'}, \quad q = q(x) = \frac{\partial f(x, y_0, y_0')}{\partial y},$$
$$\varphi(x) = f(x, y_0, y_0') - y_0''.$$

Решая линейную задачу (6) аналитически или каким-либо численным методом, найдем приближенно поправку v и примем

$$y_1 = y_0(x) + \overline{v}(x)$$

за следующее приближение.

Описанная процедура может применяться к нелинейной разностной краевой задаче, возникшей при аппроксимации задачи (1) (см. задачу 5 в § 2 гл. 7).

Задача

Численно найти наименьшее, а также третье по величине положительные числа λ_1, λ_3, при которых задача

$$y'' + (x-\lambda)y = 0, \quad 0 \leqslant \lambda \leqslant 1,$$
$$y(0) = y(1) = 0$$

имеет нетривиальные решения $y_1(x)$, $y_3(x) \neq 0$.

ГЛАВА 9
РАЗНОСТНЫЕ СХЕМЫ ДЛЯ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

Выше в связи с разностными схемами для обыкновенных дифференциальных уравнений мы определили понятия сходимости, аппроксимации и устойчивости. Мы доказали теорему о том, что если разностная краевая задача аппроксимирует дифференциальную задачу и устойчива, то при измельчении сетки решение разностной задачи сходится к решению дифференциальной задачи. В этой теореме содержится указание на способы построения сходящихся разностных схем для численного решения дифференциальных краевых задач: надо строить аппроксимирующие разностные схемы и выбирать среди них устойчивые.

Определения сходимости, аппроксимации и устойчивости и теорема о связи между этими понятиями носят общий характер. Эти понятия имеют смысл для любых функциональных уравнений. Мы иллюстрировали их примерами разностных схем для обыкновенных дифференциальных уравнений и для интегрального уравнения. Здесь мы проиллюстрируем некоторые основные способы построения разностных схем и проверки их устойчивости примерами разностных схем для уравнений с частными производными. При этом обнаружится много важных и существенно новых по сравнению со случаем обыкновенных дифференциальных уравнений обстоятельств. Главные из них — разнообразие сеток и способов аппроксимации, неустойчивость большинства взятых наудачу аппроксимирующих схем, сложность исследований устойчивости и трудности вычисления решений разностных краевых задач, требующие специальных усилий для их преодоления.

§ 1. Основные определения и их иллюстрация

1. Определение сходимости. Пусть требуется приближенно вычислить решение u дифференциальной краевой задачи

$$Lu = f, \tag{1}$$

поставленной в некоторой области D с границей Γ. Для этого следует выбрать дискретное множество точек D_h (сетку), принадлежащее $D \cup \Gamma$, ввести линейное нормированное пространство U_h функций, определенных на сетке D_h, установить соответствие между решением u и функцией $[u]_h \in U_h$, которую будем считать искомой таблицей решения u. Для приближенного отыскания таблицы $[u]_h$, которую мы условились считать точным решением задачи (1), надо на основе этой задачи составить такую систему уравнений

$$L_h u^{(h)} = f^{(h)} \tag{2}$$

относительно функции $u^{(h)}$ из U_h, чтобы имела место сходимость

$$\|[u]_h - u^{(h)}\|_{U_h} \to 0, \quad h \to 0. \tag{3}$$

Если для решения разностной краевой задачи (2) выполнено неравенство

$$\|[u]_h - u^{(h)}\|_{U_h} \leqslant ch^k, \quad c = \text{const},$$

то говорят, что *сходимость* (в смысле выбранной нами нормы) *имеет порядок k относительно h.*

Задачу построения сходящейся разностной схемы (2) разбивают на две — на построение разностной схемы (2), аппроксимирующей задачу (1) на решении u последней, и на проверку устойчивости схемы (2).

2. Определение аппроксимации. Чтобы понятие аппроксимации имело смысл, надо ввести норму в пространстве F_h, которому принадлежит правая часть $f^{(h)}$ уравнения (2). По определению разностная задача (2) аппроксимирует задачу (1) на решении u, если в равенстве

$$L_h[u]_h = f^{(h)} + \delta f^{(h)}$$

невязка $\delta f^{(h)}$, возникающая при подстановке $[u]_h$ в разностную краевую задачу (2), стремится к нулю при $h \to 0$:

$$\|\delta f^{(h)}\|_{F_h} \to 0.$$

Если

$$\|\delta f^{(h)}\|_{F_h} \leqslant ch^k,$$

где c не зависит от h, то аппроксимация имеет порядок k относительно h.

Построим, например, для задачи Коши

$$\begin{gathered}\frac{\partial u}{\partial t}-\frac{\partial u}{\partial x}=\varphi(x,t),\ -\infty<x<\infty,\ 0\leqslant t\leqslant T,\\ u(x,0)=\psi(x),\quad -\infty<x<\infty,\end{gathered}\tag{4}$$

одну из аппроксимирующих ее разностных схем. Задача (4) записывается в форме (1), если положить

$$Lu\equiv\begin{cases}\dfrac{\partial u}{\partial t}-\dfrac{\partial u}{\partial x}, & -\infty<x<\infty,\ 0\leqslant t\leqslant T,\\ u(x,0), & -\infty<x<\infty,\end{cases}$$

$$f=\begin{cases}\varphi(x,t), & -\infty<x<\infty,\ 0\leqslant t\leqslant T,\\ \psi(x), & -\infty<x<\infty.\end{cases}$$

В качестве сетки D_h используем совокупность точек пересечения прямых

$$x=mh,\quad t=p\tau,\quad m=0,\pm1,...;\quad p=0,1,...,[T/\tau],$$

где $h>0$, $\tau>0$ — некоторые числа, а $[T/\tau]$ — целая часть дроби T/τ. Будем считать, что шаг τ связан с шагом h зависимостью $\tau=rh$, где $r=\text{const}$, так что сетка D_h зависит только от одного параметра h. Искомой сеточной функцией является таблица $[u]_h=\{u(mh,p\tau)\}$ значений решения $u(x,t)$ задачи (4) в точках сетки D_h.

Перейдем к построению аппроксимирующей задачу (4) разностной схемы (2). Значения сеточной функции $u^{(h)}$ в точке $(x_m,t_p)=(mh,p\tau)$ сетки D_h будем обозначать u_m^p. Схему (2) получим, приблизив производные $\partial u/\partial t$, $\partial u/\partial x$ разностными отношениями

$$\begin{gathered}\left.\frac{\partial u}{\partial t}\right|_{x,t}\approx\frac{u(x,t+\tau)-u(x,t)}{\tau},\\ \left.\frac{\partial u}{\partial x}\right|_{x,t}\approx\frac{u(x+h,h)-u(x,t)}{h}.\end{gathered}\tag{4$'$}$$

Эта схема имеет вид

$$\begin{gathered}\frac{u_m^{p+1}-u_m^p}{\tau}-\frac{u_{m+1}^p-u_m^p}{h}=\varphi_m^p,\quad m=0,\pm1,...;\\ p=0,1,...,[T/\tau]-1,\\ u_m^0=\psi_m,\quad \varphi_m^p\equiv\varphi(mh,p\tau),\quad \psi_m=\psi(mh).\end{gathered}\tag{5}$$

Оператор L_h и правая часть $f^{(h)}$ для схемы (5) задаются соответственно равенствами

$$L_hu^{(h)}=\begin{cases}\dfrac{u_m^{p+1}-u_m^p}{h}-\dfrac{u_{m+1}^p-u_m^p}{h}, & p=0,1,...,\left[\dfrac{T}{\tau}\right]-1,\\ u_m^0,\quad m=0,\pm1,...,\end{cases}$$

$$f^{(h)}=\begin{cases}\varphi_m^p, & m=0,\pm1,...;\quad p=0,1,...,\left[\dfrac{T}{\tau}\right]-1,\\ \psi_m, & m=0,\pm1,...\end{cases}$$

Таким образом, $f^{(h)}$ — это пара сеточных функций φ_m^p, ψ_m, одна из которых задана на двумерной сетке $(x_m, t_r) = (mh, r\tau)$, $m = 0, \pm 1, ...$; $p = 0, 1, ..., [T/\tau]$, а другая — на одномерной сетке

$$(x_m, 0) = (mh, 0), \quad m = 0, \pm 1, ...$$

Разностное уравнение (5) можно разрешить относительно u_m^{p+1}, получив

$$u_m^{p+1} = (1-r)u_m^p + ru_{m+1}^p + \tau\varphi_m^p. \tag{6}$$

Итак, зная значения u_m^p $(m = 0, \pm 1, ...)$ решения $u^{(h)}$ в точках сетки при $t = p\tau$, можно вычислить значения u_m^{p+1} в точках сетки при $t = (p+1)\tau$. Поскольку значения u_m^0 при $t = 0$ заданы равенствами $u_m^0 = \psi_m$, мы можем шаг за шагом вычислить значения решения u_m^p в точках сетки на прямых $t = \tau, 2\tau, ...$, т. е. всюду на D_h.

Перейдем к выяснению порядка аппроксимации, которым обладает схема (5). За F_h можно принять линейное пространство всех пар ограниченных функций $g^{(h)} = \begin{bmatrix} \varphi_m^p \\ \psi_m \end{bmatrix}$, положив *)

$$\|g^{(h)}\|_{F_h} = \max_{m,p} |\varphi_m^p| + \max_m |\psi_m|. \tag{7}$$

Вообще говоря, норма, в которой рассматривается аппроксимация, может быть выбрана многими способами, и выбор этот небезразличен. Будем иметь в виду всюду в этом параграфе именно норму (7).

Предположим, что решение $u(x,t)$ задачи (4) имеет ограниченные вторые производные. Тогда по формуле Тейлора

$$\frac{u(x_m+h, t_p) - u(x_m, t_p)}{h} = \frac{\partial u(x_m, t_p)}{\partial x} + \frac{h}{2}\frac{\partial^2(x_m+\xi, t_p)}{\partial x^2},$$

$$\frac{u(x_m, t_p+\tau) - u(x_m, t_p)}{\tau} = \frac{\partial u(x_m, t_p)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u(x_m, t_p+\eta)}{\partial t^2},$$

где ξ, η — некоторые числа, зависящие от m, p, h и удовлетворяющие неравенствам $0 < \xi < h$, $0 < \eta < \tau$.

С помощью этих формул выражение

$$L_h[u]_h = \begin{cases} \dfrac{u(x_m, t_p+\tau) - u(x_m, t_p)}{\tau} - \dfrac{u(x_m+h, t_p) - u(x_m, t_p)}{h}, \\ u(x_m, 0) \end{cases}$$

можно переписать в виде

$$L_h[u]_h = \begin{cases} \left(\dfrac{\partial u}{\partial t} - \dfrac{\partial u}{\partial x}\right)\Big|_{x_m, t_p} + \dfrac{\tau}{2}\,\dfrac{\partial^2 u(x_m, t_p+\eta)}{\partial t^2} - \dfrac{h}{2}\,\dfrac{\partial^2 u(x_m+\xi, t_p)}{\partial x^2}, \\ u(x_m, 0) + 0, \end{cases}$$

*) Если $\max|\varphi_m^p|$ или $\max|\psi_m|$ не достигается, то имеется в виду точная верхняя грань $\sup|\varphi_m^p|$ или $\sup|\psi_m|$.

или

$$L_h[u]_h = f^{(h)} + \delta f^{(h)},$$

где

$$\delta f^{(h)} = \begin{cases} \dfrac{\tau}{2}\dfrac{\partial^2 u(x_m, t_p+\eta)}{\partial t^2} - \dfrac{h}{2}\dfrac{\partial^2 u(x_m+\xi, t_p)}{\partial x^2}, \\ 0. \end{cases}$$

Следовательно,

$$\|\delta f^{(h)}\|_{F_h} \leqslant \frac{1}{2}\left(r \sup\left|\frac{\partial^2 u}{\partial t^2}\right| + \sup\left|\frac{\partial^2 u}{\partial x^2}\right|\right)h.$$

Таким образом, рассматриваемая разностная схема (5) имеет первый порядок аппроксимации относительно h на решении $u(x,t)$, обладающем ограниченными вторыми производными.

3. Определение устойчивости. Разностная краевая задача (2) по определению *устойчива*, если существуют числа $\delta > 0$, $h_0 > 0$ такие, что при любом $h < h_0$ и любом $\varepsilon^{(h)}$ из F_h, удовлетворяющем неравенству $\|\varepsilon^{(h)}\|_{F_h} < \delta$, разностная краевая задача

$$L_h z^{(h)} = f^{(h)} + \varepsilon^{(h)}$$

имеет одно и только одно решение, причем выполняется условие

$$\|z^{(h)} - u^{(h)}\|_{U_h} \leqslant c\|\varepsilon^{(h)}\|_{F_h},$$

где c — некоторая постоянная, не зависящая от h.

В § 3 гл. 8, где введено понятие устойчивости, показано, что в случае линейного оператора L_h сформулированное определение равносильно следующему.

Разностная краевая задача (2) *устойчива*, если существует $h_0 > 0$ такое, что при $0 < h < h_0$ и любом $f^{(h)} \in F_h$ она однозначно разрешима, причем

$$\|u^{(h)}\|_{U_h} \leqslant c\,\|f^{(h)}\|_{F_h}, \tag{8}$$

где c — некоторая постоянная, не зависящая от h, $f^{(h)}$.

Свойство устойчивости можно трактовать как равномерную относительно h чувствительность решения разностной краевой задачи (2) к возмущениям $\varepsilon^{(h)}$ правой части.

Подчеркнем, что в силу приведенного определения устойчивость есть некоторое внутреннее свойство разностной краевой задачи. Оно формулируется независимо от какой-либо связи с дифференциальной краевой задачей, в частности независимо от аппроксимации или сходимости.

Теорема. *Если разностная краевая задача* (2) *аппроксимирует на решении и дифференциальную краевую задачу* (1) *и устойчива, то имеет место сходимость* (3). *При этом порядок относительно h скорости сходимости совпадает с порядком аппроксимации.*

Доказательство этой важной теоремы проведено в § 3 гл. 8.

Покажем, что разностная схема (5) при $r = \tau/h \leqslant 1$ устойчива. При этом норму $\|u^{(h)}\|_{U_h}$ определим равенством

$$\|u^{(h)}\|_{U_h} = \max_p \sup_m |u_m^p|. \qquad (9)$$

Норму $\|\cdot\|_{F_h}$ будем понимать как

$$\|f^{(h)}\|_{F_h} = \max_p \sup_m |\varphi_m^p| + \sup_m |\psi_m|.$$

Разностную задачу

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = \varphi_m^p, \quad m = 0, \pm 1, \ldots; \quad p = 0, 1, \ldots, [T/\tau] - 1, \qquad (5')$$

которая отличается от задачи (5) только тем, что φ_m^p, ψ_m — произвольные правые части, вообще говоря, не совпадающие с $\varphi(mh, p\tau)$, $\psi(mh)$, перепишем в форме

$$u_m^{p+1} = (1 - r)u_m^p + ru_{m+1}^p + \tau\varphi_m^p, \quad u_m^0 = \psi_m. \qquad (6')$$

Поскольку $r \leqslant 1$, то $1 - r \geqslant 0$. В этом случае справедлива оценка

$$|(1 - r)u_m^p + ru_{m+1}^p| \leqslant |(1 - r) + r| \max\{|u_m^p|, |u_{m+1}^p|\} = \max\{|u_m^p|, |u_{m+1}^p|\} \leqslant \sup_m |u_m^p|.$$

Используя эту оценку, выводим из (6′) неравенство

$$|u_m^{p+1}| < \sup_m |u_m^p| + \tau \sup_m |\varphi_m^p| \leqslant \sup_m |u_m^p| + \tau \sup_{m,p} |\varphi_m^p|. \qquad (6'')$$

Отметим, что в случае $\varphi_m^p \equiv 0$ из неравенства (6″) следует, что $\sup_m |u_m^p|$ не возрастает с ростом p. Отмеченное свойство разностной схемы принято называть *принципом максимума.* Для краткости будем иногда пользоваться этим названием для всего неравенства (6″).

Правая часть этого неравенства не зависит от m, так что в левой части вместо $|u_m^{p+1}|$ можно написать $\sup_m |u_m^{+1}|$, получив неравенство

$$\sup_m |u_m^{p+1}| \leqslant \sup_m |u_m^p| + \tau \sup_{p,m} |\varphi_m^p|.$$

Аналогично получаем неравенства

$$\sup_m |u_m^p| \leqslant \sup_m |u_m^{p-1}| + \tau \sup_{p,m} |\varphi_m^p|,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\sup_m |u_m^1| \leqslant \sup_m |u_m^0| + \tau \sup_{m,p} |\varphi_m^p|.$$

После почленного сложения этих неравенств и приведения подобных членов получаем

$$\sup_m |u_m^{p+1}| \leqslant \sup_m |u_m^0| + (p + 1) \sup_{p,m} |\varphi_m^p|.$$

Отсюда непосредственно следует неравенство

$$\sup_m |u_m^{p+1}| \leqslant \sup_m |\psi_m| + T \sup_{p,m} |\varphi_m^p| \leqslant (1+T)\|f^{(h)}\|_{F_h}.$$

Доказанное неравенство имеет место для всех p, так что оно останется справедливым, если вместо $\sup\limits_m |u_m^{p+1}|$ написать $\max\limits_p \sup\limits_m |u_m^p| = \|u^{(h)}\|_{U_h}$:

$$\|u^{(h)}\|_{U_h} \leqslant (1+T)\|f^{(h)}\|_{F_h}. \tag{10}$$

Неравенство (10) означает устойчивость линейной задачи (5), поскольку существование и единственность решения задачи (6′) при произвольных ограниченных φ_m^p, ψ_m имеют место. Роль постоянной c в неравенстве (8) играет здесь число $1+T$.

Не следует думать, что одна только аппроксимация дифференциальной краевой задачи (1) разностной краевой задачей (2) обеспечивает сходимость (3). Мы убедились в этом в § 3 гл. 1 с помощью специально сконструированного примера аппроксимирующей, но расходящейся разностной схемы.

В случае уравнений с частными производными непригодность наудачу взятой аппроксимирующей разностной схемы является правилом, а выбор устойчивой (и, следовательно, сходящейся) разностной схемы — постоянной заботой вычислителя.

Напомним, например, что доказательство устойчивости разностной схемы (5) мы провели в предположении, что $r \leqslant 1$. В случае $r > 1$ разностная задача (5) по-прежнему аппроксимирует задачу (4), но наше доказательство устойчивости не проходит. Покажем, что в этом случае нет сходимости решения $u^{(h)}$ разностной задачи (5) к решению $[u]_h$ дифференциальной задачи (4), а значит, не может быть и устойчивости, так как аппроксимация и устойчивость влекли бы за собой сходимость.

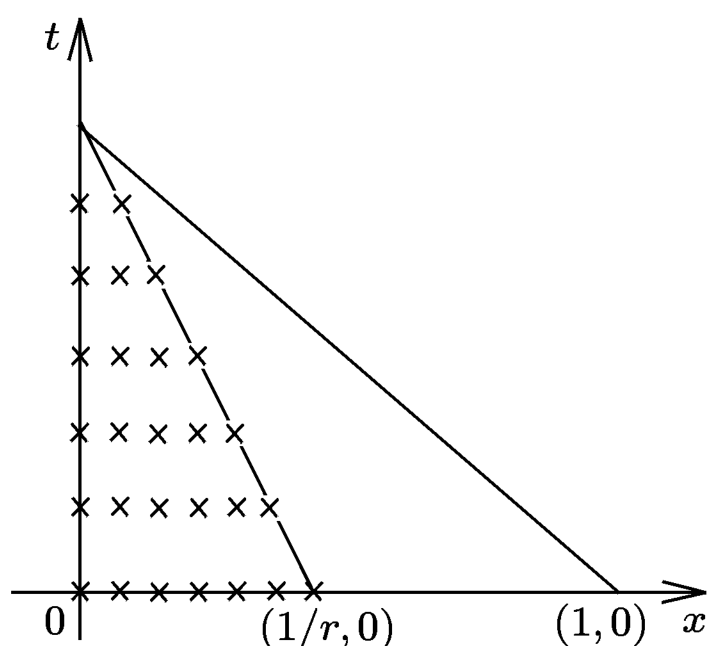


Рис. 16

4. Условие Куранта, Фридрихса и Леви. Покажем, что в случае $\tau/h > 1$ разностная схема (5) не может оказаться сходящейся при произвольной функции $\psi(x)$. Пусть для определенности $\varphi(x,t) \equiv 0$, так что $\varphi(mh, p\tau) \equiv 0$; пусть, далее, $T = 1$. Шаг h будем выбирать так, чтобы точка $(0,1)$ на плоскости Oxt принадлежала сетке, т. е. чтобы число

$$N = 1/\tau = 1/(rh)$$

было целым (рис. 16). В силу разностного уравнения имеем

$$u_m^{p+1} = (1-r)u_m^p + ru_{m+1}^p.$$

Значение u_0^{p+1} решения $u^{(h)}$ в точке $(0,1)$ сетки выражается через значения u_0^p, u_1^p решения в точках $(0, 1-\tau)$, $(h, 1-\tau)$ сетки. Значения u_0^p, u_1^p выражаются через значения u_0^{p-1}, u_1^{p-1}, u_2^{p-1} решения в трех точках сетки: $(0, 1-2\tau)$, $(h, 1-2\tau)$, $(2h, 1-2\tau)$. Значения решения u_0^{p-1}, u_1^{p-1}, u_2^{p-1} в свою очередь выражаются через значения решения в четырех точках: $(0, 1-3\tau)$, $(h, 1-3\tau)$, $(2h, 1-3\tau)$, $(3h, 1-3\tau)$, и т. д. В конечном счете значение u_0^{p+1} выражается через значения $u_m^0 = \psi_m$ решения в точках сетки $(0,0)$, $(h,0)$, $(2h,0)$, ..., $(hT/\tau, 0)$. Все эти точки лежат на отрезке

$$0 \leqslant x \leqslant h/\tau = 1/r$$

прямой $t = 0$ (см. рис. 16), где задано начальное условие

$$u(x,0) = \psi(x)$$

для дифференциального уравнения.

Таким образом, решение разностного уравнения в точке $(0,1)$ сетки не зависит от значений функции $\psi(x)$ в точках x, лежащих вне отрезка $0 \leqslant x \leqslant 1/r$. Но решением задачи

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad -\infty < x < \infty, \quad t > 0,$$
$$u(x,0) = \psi(x), \quad -\infty < x < \infty,$$

как легко проверить, является функция

$$u(x,t) \equiv \psi(x+t).$$

Она постоянна на каждой характеристике $x + t = \text{const}$ и, в частности, на прямой $x + t = 1$, которая проходит через точки $(0,1)$, $(1,0)$ (см. рис. 16) и в точке $(0,1)$ принимает значение $\psi(1)$. Отсюда видно, что в случае $r > 1$ сходимости, вообще говоря, быть не может.

Действительно, в этом случае отрезок оси абсцисс $0 \leqslant x \leqslant 1/r < 1$ не содержит точку $(1,0)$. Если бы при какой-нибудь функции $\psi(x)$ сходимость (случайно) имела место, то, не меняя значения $\psi(x)$ на отрезке $0 \leqslant x \leqslant 1/r$ и не меняя, таким образом, значения решения разностного уравнения в точке $(0,1)$, мы могли бы нарушить сходимость, изменив $\varphi(x)$ в точке $x = 1$ и ее окрестности, что отразилось бы на значении $u(0,1)$ решения дифференциального уравнения. Изменение $\psi(x)$ в точке $x = 1$ и ее окрестности можно внести так, чтобы не нарушить существования вторых производных функции $\psi(x)$ и решения $u(x,t) = \psi(x+t)$, так что аппроксимация на решении $u(x,t)$ будет иметь место. В этих условиях из устойчивости схемы (5) вытекала бы сходимость. Но поскольку при $r > 1$ нет сходимости, то нет и устойчивости.

Соображение, которое мы использовали для доказательства непригодности схемы (5) в случае $r = \tau/h > 1$, носит общий характер. Оно состоит в следующем.

Допустим, что в постановке исходной задачи участвует некоторая функция ψ (см., например, задачу (4)). Выберем произвольную точку P, принадлежащую области определения решения u. Пусть значение $u(P)$ зависит от значения функции ψ в точках некоторого множества $G_\psi = G_\psi(P)$, принадлежащего области определения функции ψ. Это значит, что, изменяя значения ψ в малой окрестности любой точки Q из области $G_\psi(P)$, можно вызвать изменение значения решения $u(P)$.

Допустим, что для вычисления решения u используется некоторая разностная схема $L_h u^{(h)} = f^{(h)}$, причем значение решения $u^{(h)}$ в точке сетки $P^{(h)}$, ближайшей к P, полностью определяется значениями функции ψ на некотором множестве $G_\psi^{(h)}(P)$.

Условие Куранта, Фридрихса и Леви состоит в следующем. *Для того чтобы имела место сходимость $u^{(h)} \to u$ при $h \to 0$, разностная схема необходимо должна быть устроена так, чтобы при $h \to 0$ в произвольной окрестности любой точки области $G_\psi(P)$ при достаточно малом h имелась точка множества $G_\psi^{(h)} = G_\psi^{(h)}(P)$.*

Объясним, почему в случае невыполнения сформулированного условия Куранта, Фридрихса и Леви сходимости ожидать не приходится. Пусть это условие не выполнено, так что в некоторой фиксированной окрестности некоторой точки Q из области $G_\psi(P)$ при всех достаточно малых h нет точек из множества $G_\psi^{(h)}(P)$. Если сходимость $u^{(h)} \to u$ при данной функции ψ имеет (случайно) место, то изменим ψ в указанной окрестности точки Q так, чтобы изменилось значение u, оставляя вне этой окрестности функцию ψ неизменной. Сходимость $u^{(h)} \to u$ при новой функции ψ уже не может иметь места: значение $u(P)$ изменилось, в то время как значения $u^{(h)}$ в точке сетки $P^{(h)}$, ближайшей к P, остались при малых h неизменными, поскольку функция ψ в точках множеств $G_\psi^{(h)} = G_\psi^{(h)}(P)$ осталась неизменной.

Условию Куранта, Фридрихса и Леви нетрудно придать форму теоремы, а проведенные рассуждения превратить в ее доказательство, но мы не будем этого делать.

Подчеркнем еще раз, что в случае, если разностная схема $L_h u^{(h)} = f^{(h)}$ аппроксимирует исходную задачу на решении u, то условие Куранта, Фридрихса и Леви является не только необходимым условием сходимости, но также и необходимым условием устойчивости схемы $L_h u^{(h)} = f^{(h)}$.

В случае невыполнения этого условия устойчивости быть не может, так как аппроксимация и устойчивость влекли бы за собой сходимость в силу теоремы, доказанной выше.

5. Механизм неустойчивости. Проведенное нами доказательство неустойчивости схемы (5) основано на использовании условия

Куранта, Фридрихса и Леви, необходимого для сходимости и устойчивости. Оно носит, таким образом, косвенный характер. Интересно проследить непосредственно, как сказывается неустойчивость при $r > 1$ разностной схемы (5) на чувствительности решения $u^{(h)}$ к погрешностям в задании $f^{(h)}$. Ведь именно равномерная относительно h чувствительность решения к погрешностям при задании $f^{(h)}$ и определена выше как устойчивость.

Допустим, что при всех h выполняются тождества $\varphi_m^p \equiv 0$, $\psi_m \equiv 0$, так что

$$f^{(h)} = \left\{ \begin{array}{c} \varphi_m^p \\ \psi_m \end{array} \right\} = 0,$$

и решение $u^{(h)} = \{u_m^p\}$ задачи (5) есть тождественный нуль: $u_m^p \equiv 0$. Допустим, далее, что при задании начальных данных допущена погрешность и вместо $\psi_m = 0$ задано $\widetilde{\varphi}_m = (-1)^m \varepsilon$ $(\varepsilon = \text{const})$, так что вместо $f^{(h)} = 0$ задано

$$f^{(h)} = \left\{ \begin{array}{c} 0 \\ \widetilde{\psi}_m \end{array} \right\}, \quad \|f^{(h)}\|_{F_h} = \varepsilon.$$

Будем обозначать получающееся при этом решение через $\widetilde{u}^{(h)}$. В силу уравнений

$$\widetilde{u}_m^{p+1} = (1-r)\widetilde{u}_m^p + r\widetilde{u}_{m+1}^p, \quad \widetilde{u}_m^0 = (-1)^m \varepsilon$$

для $\widetilde{u}_m^1$ получим $\widetilde{u}_m^1 = (1-r)\widetilde{u}_m^0 + r\widetilde{u}_{m+1}^0 = (1-2r)\widetilde{u}_m^0$.

Мы видим, что допущенная при $p = 0$ погрешность умножилась на число $1 - 2r$. При переходе к $\widetilde{u}_m^0$ получим

$$\widetilde{u}_m^2 = (1-2r)^2 \widetilde{u}_m^0.$$

Вообще,

$$\widetilde{u}_m^p = (1-2r)^p \widetilde{u}_m^0.$$

При $r > 1$ будем иметь $1 - 2r < -1$, так что погрешность

$$\widetilde{u}_m^0 = (-1)^m \varepsilon$$

при переходе от одного слоя $t = p\tau$ сетки к следующему умножается на отрицательное число, превосходящее единицу по модулю. При $p = [T/\tau]$

$$|\widetilde{u}_m^p| = |1-2r|^{[T/\tau]} |\widetilde{u}_m^0|.$$

Отсюда

$$\|\widetilde{u}^{(h)}\|_{U_h} = \sup_{m,r} |\widetilde{u}_m^p| = |1-2r|^{[T/(rh)]} \sup_m |\widetilde{\psi}_m| = |1-2r|^{[T/(rh)]} \|f^{(h)}\|_{F_h}.$$

При фиксированном T первоначально допущенная в начальных данных погрешность $(-1)^m \varepsilon$ увеличивается в очень быстро возрастающее при $h \to 0$ число раз, равное $|1-2r|^{[T/(\tau h)]}$.

6. Об экономичности разностной схемы. Остановимся теперь кратко на критике принятого нами способа оценки качества аппроксимации сравнением величины нормы невязки $\|\delta f^{(h)}\|_{F_h}$ с той или иной степенью h. Как мы знаем, для устойчивых схем порядок аппроксимации совпадает с порядком погрешности $\|[u]_h - u^{(h)}\|_{U_h}$ в решении. Качество схемы естественно оценивать по количеству вычислительной работы, необходимой для получения заданной точности. Количество же работы, вообще говоря, пропорционально числу точек N использованной разностной сетки. Для обыкновенных дифференциальных уравнений N обратно пропорционально шагу h. Поэтому, когда мы говорим, что ε — норма погрешности, где $\varepsilon \approx h^k$, мы тем самым утверждаем, что $\varepsilon = N^{-k}$, т. е. что уменьшение погрешности вдвое требует увеличения работы в $\sqrt[k]{2}$ раз. Таким образом, в случае обыкновенных разностных уравнений порядок аппроксимации относительно h характеризует объем работы.

Для уравнений с частными производными дело обстоит уже не так. В рассмотренном нами примере задачи с двумя переменными x, t сетка задается двумя шагами h, τ. Число N точек сетки, помещающихся в ограниченной области на плоскости Oxt, имеет порядок $1/(\tau h)$. Это число также может применяться для оценки количества работы, затрачиваемой при решении разностных уравнений. Пусть $\tau = rh$. В этом случае $N \approx h^{-2}$, и утверждение, что $\varepsilon = h^k$, эквивалентно утверждению $\varepsilon \approx N^{-k/2}$. Если $\tau = rh^2$, то $N \approx h^{-3}$, и утверждение $\varepsilon \approx h^k$ эквивалентно тому, что $\varepsilon \approx N^{-1/3}$.

Мы видим, что в случае уравнений с частными производными порядок погрешности естественнее было бы измерять не в степенях h, а в степенях N^{-1}. Мы все же остановимся на описанном выше способе оценки аппроксимации степенями, так как это удобнее при проведении выкладок. Читатель, однако, должен при оценке качества разностных схем иметь в виду отмеченное обстоятельство.

Надо еще заметить, что утверждение о пропорциональности вычислительной работы числу N точек сетки тоже не всегда является верным. Можно привести примеры разностных схем, для вычисления решения по которым требуется произвести $\asymp N^{1+q}$ арифметических операций, где q равно $1/2$, 1 или даже 2. С этим приходится встречаться при решении разностных краевых задач, аппроксимирующих эллиптические уравнения, или при решении задач в случае трех и более независимых переменных (например, $u(t,x,y)$).

При реальных расчетах на компьютере для сравнительной оценки используемых алгоритмов за меру качества схемы обычно естественно принять машинное время. Машинное время не обязательно пропорционально числу арифметических действий. Играют роль (иногда превалирующую) затраты времени на пересылку информации из одного блока машинной памяти в другой. Может играть роль время,

расходуемое на логические операции.

7. Библиографическая справка. Понятие устойчивости разностных схем относительно ошибок округления при задании начальных данных впервые описано Дж. фон Нейманом и Р.Д. Рихтмайером в 1950 г. (см.: Механика.—1951.—Вып. 1) в работе, посвященной расчету газодинамических скачков. Первая система определений устойчивости и аппроксимации, при которой сходимость является следствием аппроксимации и устойчивости, была предложена В.С. Рябеньким (ДАН СССР.—1952.—Т. 86, № 6) в случае разностных аналогов задачи Коши для систем уравнений с частными производными.

Принятая в нашей книге система основных определений и теорема о том, что из аппроксимации и устойчивости следует сходимость, близки к предложенным А.Ф. Филипповым (ДАН СССР.—1955.—Т. 100, № 6). См. также [17]. Отличие состоит главным образом в том, что мы используем более универсальное, чем использовал А.Ф. Филиппов, определение аппроксимации.

Существуют другие естественные системы определений основных понятий, при которых аппроксимация и устойчивость обеспечивают сходимость. Среди них наиболее известна система определений П.Д. Лакса. В теории Лакса рассматриваются разностные схемы для нестационарных задач, причем предполагается, что эти разностные схемы действуют не в пространстве сеточных функций, а в том же функциональном пространстве, что и дифференциальное уравнение. При этом предположении доказывается, что для аппроксимирующей разностной схемы устойчивость и сходимость имеют место одновременно.

В последующие годы А.А. Самарский предложил и развил в соавторстве с А.В. Гулиным теорию устойчивости, применимую к весьма широкому классу разностных схем [18].

Следует сказать, что в работе 1928 г. Р. Куранта, К. Фридрихса и Г. Леви (см.: УМН.—1940.—Т. 8) и во многих других работах, где метод конечных разностей используется для доказательства существования решений дифференциальных уравнений, устанавливаются неравенства, которые в современной терминологии можно истолковать как устойчивость в тех или иных нормах. Однако понятие устойчивости возникло в связи с использованием разностных схем для приближенного вычисления решений в предположении, что эти решения существуют. Поэтому устойчивость изучается обычно в более слабых нормах, чем это нужно для доказательства существования.

Отметим, что впервые метод конечных разностей был использован для доказательства существования решений уравнений с частными производными в 1924 г. Л.А. Люстерником (см.: УМН.—1940.—Т. 8), который рассматривал уравнение Лапласа.

Задачи

1. Для задачи Коши (4) исследовать следующую разностную схему:

$$\frac{u_m^{p+1}-u_m^p}{\tau}-\frac{u_m^p-u_{m-1}^p}{h}=\varphi_m^p, \quad m=0,\pm1,\ldots,$$
$$u_m^0=\psi_m, \quad p=0,1,\ldots,[T/\tau]-1,$$

где $\tau=rh$, $r=\mathrm{const}$.

а) Выписать оператор L_h и правую часть $f^{(h)}$, возникающие при записи этой схемы в виде $L_hu^{(h)}=f^{(h)}$.

б) Изобразить взаимное расположение трех точек сетки (шаблон), значения $u^{(h)}$ в которых связывает разностное уравнение при фиксированных m,p.

в) Показать, что разностная схема аппроксимирует дифференциальную задачу с первым относительно h порядком на решении $u(x,t)$, имеющим ограниченные вторые производные.

г) Выяснить, устойчива ли исследуемая разностная схема при каком-либо выборе r.

2. Для задачи Коши

$$u_t+u_x=\varphi(x,t), \quad -\infty<x<\infty,$$
$$u(x,0)=\psi(x), \quad 0<t<T,$$

исследовать по предложенному в задаче 1 плану каждую из следующих разностных схем:

$$\frac{u_m^{p+1}-u_m^p}{\tau}+\frac{u_m^p-u_{m-1}^p}{h}=\varphi_m^p, \quad u_m^0=\psi_m;$$
$$\frac{u_m^{p+1}-u_m^p}{\tau}+\frac{u_{m+1}^p-u_m^p}{h}=\varphi_m^p, \quad u_m^0=\psi_m.$$

3. Для задачи Коши для уравнения теплопроводности

$$\begin{gathered}\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}=\varphi(x,t), \quad -\infty<x<\infty, \quad 0\leqslant t\leqslant T,\\ u(x,0)=\psi(x), \quad -\infty<x<\infty,\end{gathered} \tag{11}$$

рассмотреть разностную схему

$$\begin{gathered}\frac{u_m^{p+1}-u_m^p}{\tau}-\frac{u_{m+1}^p-2u_m^p+u_{m-1}^p}{h^2}=\varphi(mh,p\tau),\\ m=0,\pm1,\ldots; \quad p=0,1,\ldots,[T/\tau]-1,\\ u_m^0=\psi(mh), \quad m=0,\pm1,\ldots\end{gathered} \tag{12}$$

Нормы в U_h, F_h понимаем в смысле (9), (7).

а) Проверить, что при $\tau/h^2=r=\mathrm{const}$ разностная схема (12) аппроксимирует задачу (11) с порядком $O(h^2)$.

б*) Показать, что в случае $r\leqslant 1/2$, $\varphi(x,t)\equiv 0$ справедлив следующий принцип максимума:

$$\sup_m |u_m^{p+1}|\leqslant \sup_m |u_m^p|.$$

в*) Опираясь на принцип максимума, доказать, что в случае $r\leqslant 1/2$ разностная схема (12) устойчива.

4. Решение задачи Коши (11) для уравнения теплопроводности имеет вид

$$u(x,t) = \frac{1}{2\sqrt{\pi}} \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{t}} \psi(s) \exp\left\{ -\frac{(x-\xi)^2}{4t} \right\} d\xi.$$

Может ли оказаться устойчивой и сходящейся "явная" разностная схема (12), аппроксимирующая эту задачу, если $\tau = h,\ h \to 0$?

Сопоставить области влияния для дифференциальной и разностной задач и воспользоваться условием Куранта.

5. Система уравнений акустики

$$\frac{\partial v}{\partial t} = \frac{\partial w}{\partial x}, \quad \frac{\partial w}{\partial t} = \frac{\partial v}{\partial x}, \quad -\infty < x < \infty, \quad 0 \leqslant t \leqslant T,$$
$$v(x,0) = \varphi(x), \quad w(x,0) = \psi(x),$$

имеет решение вида

$$v(x,t) = \frac{\varphi(x-t) - \psi(x-t) + \varphi(x+t) + \psi(x+t)}{2},$$
$$w(x,t) = \frac{-\varphi(x-t) + \psi(x-t) + \varphi(x+t) + \psi(x+t)}{2}.$$

Может ли оказаться сходящейся разностная схема вида

$$\frac{v_m^{p+1} - v_m^p}{\tau} + \frac{w_{m+1}^p - w_m^p}{h} = 0, \quad m = 0, \pm 1, ..., p \geqslant 0,$$
$$\frac{w_m^{p+1} - w_m^p}{\tau} + \frac{v_{m+1}^p - v_m^p}{h} = 0, \quad v_m^0 = \varphi(mh), \quad w_m^0 = \psi(mh)?$$

Сопоставить области влияния начальных данных для разностной и дифференциальной задач.

6*. Задача Коши

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x}, \quad t > 0, \quad -\infty < x < \infty,$$
$$u(x,0) = e^{i\alpha x}, \quad 0 < \alpha < 2\pi, \quad \alpha = \mathrm{const},$$

имеет решение

$$u(x,t) = e^{i\alpha t} e^{i\alpha x}.$$

Соответствующая разностная схема

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad p = 0, 1, ..., \quad m = 0, \pm 1, ...,$$
$$u_m^0 = e^{i\alpha(mh)}$$

имеет решение

$$u_m^p = (1 - r + re^{i\alpha h})^p e^{i\alpha(mh)},$$

которое при $p = t/\tau$, $m = x/h$ стремится к решению дифференциальной задачи при $h \to 0$, каково бы ни было фиксированное $r = \tau/h$. Между тем при $r > 1$ разностная схема не удовлетворяет условию Куранта, Фридрихса и Леви, необходимому для сходимости. Объясните кажущийся парадокс.

§ 2. Некоторые приемы построения аппроксимирующих разностных схем

1. Замена производных разностными отношениями. Простейший прием построения разностных краевых задач, аппроксимирующих дифференциальные, состоит в замене производных соответствующими разностными отношениями. Приведем несколько примеров разностных схем, полученных таким способом. В этих примерах будут использованы приближенные формулы

$$\begin{gathered}\frac{df(z)}{dz} \approx \frac{f(z+\Delta z)-f(z)}{\Delta z}, \quad \frac{df(z)}{dz} \approx \frac{f(z)-f(z-\Delta z)}{\Delta z}, \\ \frac{df(z)}{dz} \approx \frac{f(z+\Delta z)-f(z-\Delta z)}{2\Delta z}, \\ \frac{d^2 f(z)}{dz^2} \approx \frac{f(z+\Delta z)-2f(z)+f(z-\Delta z)}{(\Delta z)^2}.\end{gathered} \tag{1}$$

Предполагая функцию $f(z)$ имеющей достаточное число ограниченных производных, можно выписать выражение для остаточных членов этих формул. По формуле Тейлора

$$\begin{gathered}f(z+\Delta z) = f(z) + \Delta z f'(z) + \frac{(\Delta z)^2}{2!} f''(z) + \\ + \frac{(\Delta z)^3}{3!} f'''(z) + \frac{(\Delta z)^4}{4!} f^{(4)}(z) + o\left[(\Delta z)^4\right], \\ f(z-\Delta z) = f(z) - \Delta z f'(z) + \frac{(\Delta z)^2}{2!} f''(z) - \\ - \frac{(\Delta z)^3}{3!} f'''(z) + \frac{(\Delta z)^4}{4!} f^{(4)}(z) + o\left[(\Delta z)^4\right].\end{gathered} \tag{2}$$

Используя разложения (2), можно получить выражения для остаточных членов приближенных формул (1). Именно, справедливы равенства

$$\begin{aligned}\frac{f(z+\Delta z)-f(z)}{\Delta z} &= f'(z) + \left[\frac{\Delta z}{2} f''(z) + o(\Delta z)\right], \\ \frac{f(z)-f(z-\Delta z)}{\Delta z} &= f'(z) + \left[-\frac{\Delta z}{2} f''(z) + o(\Delta z)\right], \\ \frac{f(z+\Delta z)-2f(z)+f(z-\Delta z)}{(\Delta z)^2} &= f''(z) + \left[\frac{(\Delta z)^2}{12} f^{(4)}(z) + o((\Delta z)^2)\right].\end{aligned} \tag{3}$$

Остаточные члены приближенных формул (1) входят в соответствующие равенства (3) в виде выражений в квадратных скобках.

Очевидно, что формулы (1) и выражения остаточных членов, выписанные в формулах (3), можно использовать и при замене частных

производных разностными отношениями. Например,

$$\frac{\partial u(x,t)}{\partial t} \approx \frac{u(x,t+\Delta t) - u(x,t)}{\Delta t},$$

причем

$$\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} = \frac{\partial u(x,t)}{\partial t} + \left[\frac{\Delta t}{2}\,\frac{\partial^2 u(x,t)}{\partial t^2} + o(\Delta t)\right].$$

Точно так же справедлива формула

$$\frac{\partial u(x,t)}{\partial x} \approx \frac{u(x+\Delta x,t) - u(x,t)}{\Delta x},$$

и при этом

$$\frac{u(x+\Delta x,t) - u(x,t)}{\Delta x} = \frac{\partial u(x,t)}{\partial x} + \left[\frac{\Delta x}{2}\,\frac{\partial^2 u(x,t)}{\partial x^2} + o(\Delta x)\right]$$

и т. д.

Пример 1. Вернемся к задаче Коши (4) из § 5:

$$\begin{gathered} \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \varphi(x,t), \quad -\infty < x < \infty, \quad 0 \leqslant t \leqslant T, \\ u(x,0) = \psi(x), \quad -\infty < x < \infty. \end{gathered} \tag{4}$$

Для аппроксимации этой задачи Коши построим три схемы. Во всех этих схемах используем сетку D_h, образованную точками пересечения прямых $x = mh$, $t = p\tau$, попавшими в полосу $0 \leqslant t \leqslant T$. Значения τ, h будем считать связанными соотношением $\tau = rh$, где r — некоторая положительная постоянная. Простейшая из этих схем имеет вид

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_{m+1}^p - u_m^p}{h} = \varphi(mh, p\tau), \\ u_m^0 = \psi(mh) \end{cases} \tag{5}$$

и получается при замене производных $\partial u/\partial t$, $\partial u/\partial x$ по приближенным формулам

$$\begin{gathered} \frac{\partial u(x,t)}{\partial t} \approx \frac{u(x,t+\tau) - u(x,t)}{\tau}, \\ \frac{\partial u(x,t)}{\partial x} \approx \frac{u(x+h,t) - u(x,t)}{h}. \end{gathered}$$

Невязка $\delta f^{(h)}$, возникающая при подстановке решения $[u]_h$ дифференциальной задачи в левую часть разностной задачи

$$L_h[u]^h = f^{(h)} + \delta f^{(h)},$$

выражается формулой

$$\delta f^{(h)} = \begin{cases} \left(\dfrac{\tau}{2}\,u_{tt} - \dfrac{h}{2}\,u_{xx}\right)_m^p + O(\tau + h), \\ 0. \end{cases}$$

За норму элемента $f^{(h)}$ пространства F_h примем в этом параграфе максимум всех компонент элемента $f^{(h)} \in F_h$. Тогда очевидно, что

$$\|\delta f^{(h)}\|_{F_h} = O(\tau + h) = O(rh + h) = O(h),$$

и порядок аппроксимации получается первый.

Вторая схема получается при использовании другой формулы при замене $\partial u(x,t)/\partial x$:

$$\frac{\partial u(x,t)}{\partial x} \approx \frac{u(x,t) - u(x-h,t)}{h}.$$

Она имеет вид

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_m^p - u_{m-1}^p}{h} = \varphi(mh, p\tau), \\ u_m^0 = \psi_m. \end{cases}$$

Порядок аппроксимации снова получается первый.

Вторая схема, казалось бы, совсем несущественно отличается от первой. В действительности, однако, вторая схема непригодна для счета: для нее при любом $r = \tau/h$ не выполнено условие Куранта.

Для облегчения запоминания разностной схемы ее обычно принято сопоставлять с картинкой, на которой изображено взаимное расположение точек сетки (шаблон), значения в которых связывает разностное уравнение при некоторых фиксированных значениях m, p. Для двух рассмотренных схем эти картинки изображены на рис. 17.

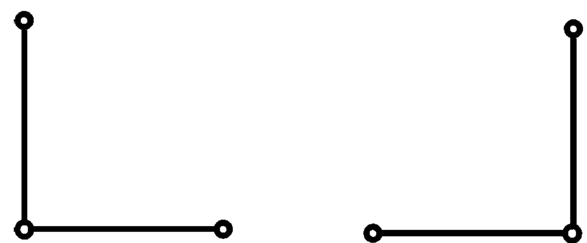


Рис. 17

Пример 2. Приведем две разностные схемы, аппроксимирующие задачу Коши для уравнения теплопроводности:

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \varphi(x,t), \quad -\infty < x < \infty, \quad 0 < t < T,$$
$$u(x,0) = \psi(x), \quad -\infty < x < \infty.$$

Простейшая из них:

$$L_h^{(1)} u^{(h)} = \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \varphi(mh, p\tau), \\ u_m^0 = \psi(mh), \end{cases}$$

$$f^{(h)} = \begin{cases} \varphi(mh, p\tau), \\ \psi(mh) \end{cases}$$

получается при замене производных u_t, u_{xx} разностными отношениями по формулам

$$u_t(x,t) \approx \frac{u(x,t+\tau) - u(x,t)}{\tau},$$
$$u_{xx}(x,t) \approx \frac{u(x-h,t) - 2u(x,t) + u(x-h,t)}{h^2}.$$

Если для замены $u_{xx}(x,t)$ использовать другую формулу:

$$u_{xx}(x,t) \approx \frac{u(x+h,t+\tau) - 2u(x,t+\tau) + u(x-h,t+\tau)}{h^2},$$

придем к другой схеме для того же уравнения:

$$L_h^{(2)} u^{(h)} = \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_{m+1}^{p+1} - 2u_m^{p+1} + u_{m-1}^{p+1}}{h^2} = \varphi(mh, p\tau), \\ u_m^0 = \psi(mh). \end{cases}$$

Чтобы различить операторы L_h этих двух схем, мы снабдили их номерами и написали $L_h^{(1)} u^{(h)} = f^{(h)}$, $L_h^{(2)} u^{(h)} = f^{(h)}$. Шаблоны, соответствующие этим разностным схемам, изображены на рис. 18.

Эти схемы существенно различаются. Вычисление решения по первой из них не представляет труда и проводится по явной формуле

$$u_m^{p+1} = (1-2r)u_m^p + r(u_{m+1}^p + u_{m-1}^p) + \tau\varphi(mh, p\tau),$$

где $r = \tau/h^2$. Эта формула получена из разностного уравнения в результате решения его относительно u_m^{p+1}. Зная значения решения



Рис. 18

u_m^p $(m = 0, \pm 1, ...)$ на слое $t = t_p = p\tau$ сетки, мы можем вычислять его значения u_m^{p+1} на следующем слое: $t = t_{p+1} = (p+1)\tau$.

Вторая схема $L_h^{(2)} u^{(h)} = f^{(h)}$ лишена этого удобного свойства. В этом случае разностное уравнение, выписанное при фиксированных m, p нельзя разрешить относительно u_m^{p+1}, выразив это значение через известные значения u_{m-1}^p, u_m^p, u_{m+1}^p с предыдущего слоя. Дело в том, что в это уравнение входит не только неизвестное значение u_m^{p+1}, но также и неизвестные u_{m+1}^{p+1}, u_{m-1}^{p+1}. Поэтому для определения u_m^{p+1} $(m = 0, \pm 1, ...)$ придется решать разностное уравнение относительно сеточной функции u_m^{p+1} аргумента m. Тем не менее в дальнейшем будет показано, что схема $L_h^{(2)} u^{(h)} = f^{(h)}$, как правило, удобнее схемы $L_h^{(1)} u^{(h)} = f^{(h)}$.

При $\tau = rh^2$ $(r = \text{const})$ обе схемы имеют второй порядок аппроксимации относительно h. Вычислим невязку $\delta f^{(h)}$ и оценим порядок аппроксимации для второй из этих схем. Пользуясь формулами (3),

можно написать

$$L_h^{(2)}[u]_h = \begin{cases} (u_t - u_{xx})\Big|_{\substack{x=mh \\ t=t_{p+1}}} - \dfrac{\tau}{2}\, u_{tt}(x, t_{p+1}) - \\ \qquad\qquad\qquad\qquad - \dfrac{h^2}{12}\, \dfrac{\partial^4 u(x, t_{p+1})}{\partial x^4} + O(\tau + h^2), \\ u(mh, 0). \end{cases}$$

Отсюда с учетом $\tau = rh^2$ можно написать

$$L_h^{(2)}[u]_h \equiv \begin{cases} \varphi(mh, (p+1)\tau) + O(h^2), \\ \psi(mh) + 0, \end{cases}$$

$$\delta f^{(h)} = \begin{cases} \varphi(x_m, t_p) - \varphi(x_m, t_{p+1}) + O(h^2), \\ 0. \end{cases}$$

Но

$$\varphi(x_m, t_{p+1}) = \varphi(x_m, t_p) + [\varphi(x_m), t_{p+1} - \varphi(x_m, t_p)] = \\ = \varphi(x_m, t_p) + O(\tau) = \varphi(x_m, t_p) + O(h^2).$$

Поэтому

$$\|\delta f^{(h)}\|_{F_h} = O(h^2).$$

Пример 3. Рассмотрим простейшую разностную схему, аппроксимирующую задачу Дирихле для уравнения Пуассона в квадрате D $(0 < x < 1,\ 0 < y < 1)$ с границей Γ (см. рис. 8):

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \varphi(x, y), \quad (x, y) \in D,$$

$$u|_\Gamma = \psi(x, y), \quad (x, y) \in \Gamma.$$

Построим сетку D_h, отнеся к ней те точки $(x_m, y_n) = (mh, nh)$, которые попали внутрь квадрата или на его границу. Шаг h будем считать выбранным так, чтобы число $1/h$ было целым.

Разностную схему $L_h u^{(h)} = f^{(h)}$ зададим равенствами

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{m+1,n} - 2u_{mn} + u_{m-1,n}}{h^2} + \dfrac{u_{m,n+1} - 2u_{mn} + u_{m,n-1}}{h^2} = \\ \qquad\qquad\qquad = \varphi(mh, nh), \quad (mh, nh) \in D_h, \\ u_{mn} = \psi_{mn}, \quad (mh, nh) \in \Gamma. \end{cases}$$

Здесь

$$f^{(h)} = \begin{cases} \varphi(mh, nh), & (mh, nh) \in D, \\ \psi(mh, nh), & (mh, nh) \in \Gamma. \end{cases}$$

Невязка $\delta f^{(h)}$ имеет в силу формул (3) вид

$$\delta f^{(h)} = \begin{cases} \dfrac{h^2}{12}\left(\dfrac{\partial^4 u}{\partial x^4} + \dfrac{\partial^4 u}{\partial y^4}\right) + O(h^2), \\ 0, \end{cases}$$

так что аппроксимация имеет второй порядок. Пятиточечный шаблон, отвечающий использованному разностному уравнению, изображен на рис. 19.

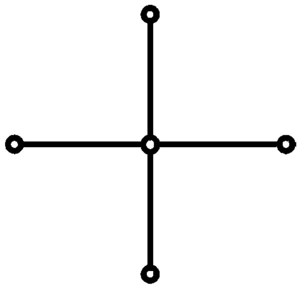


Рис. 19

Разностные схемы, построенные выше, получались путем замены каждой производной в дифференциальном уравнении тем или иным разностным отношением.

2. Метод неопределенных коэффициентов. Более общий способ построения разностных схем состоит в том, что приближается не каждая производная в отдельности, а сразу весь дифференциальный оператор. Разъясним этот способ на примерах разностных схем для задачи Коши (4). Сначала рассмотрим схему первого порядка аппроксимации (5). Она связывает значения искомой функции в трех точках, изображенных на рис. 17 слева. Разностное уравнение

$$\Lambda_h u^{(h)} = \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = \varphi(mh, nh),$$

используемое в этой схеме, имеет вид

$$\Lambda_h u^{(h)} \equiv a^0 u_m^{p+1} + a_0 u_m^p + a_1 u_{m+1}^p = \varphi(mh, p\tau).$$

Забудем на время, что нам уже известна разностная схема (5), для которой

$$a^0 = \frac{1}{\tau}, \quad a_0 = \frac{1}{h} - \frac{1}{\tau}, \quad a_1 = -\frac{1}{h},$$

и, считая эти коэффициенты неопределенными, постараемся подобрать их так, чтобы имело место равенство

$$\Lambda_h [u]_h \Big|_{\substack{x=mh \\ t=p\tau}} = \left(\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} \right)_{\substack{x=mh \\ t=p\tau}} + O(h),$$

или

$$\Lambda_h [u]_h \Big|_{\substack{x=mh \\ t=p\tau}} = \Lambda u \Big|_{\substack{x=mh \\ t=p\tau}} + O(h), \tag{6}$$

где

$$\Lambda u = \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x}. \tag{7}$$

Воспользуемся формулой Тейлора:

$$u(mh, (p+1)\tau) = u(mh, p\tau) + \tau \frac{\partial u(mh, p\tau)}{\partial t} + O(\tau^2),$$

$$u((m+1)h, p\tau) = u(mh, p\tau) + h \frac{\partial u(mh, p\tau)}{\partial x} + O(h^2).$$

Подставив эти выражения в правую часть равенства

$$\Lambda_h[u]_h\Big|_{\substack{x=mh\\t=p\tau}} \equiv a^0u(mh,(p+1)\tau)+a_0u(mh,p\tau)+a_1u((m+1)h,p\tau),$$

получим

$$\Lambda_h[u]_h\Big|_{\substack{x=mh\\t=p\tau}} = (a^0+a_0+a_1)\,u(mh,p\tau)+$$
$$+a^0\tau\,\frac{\partial u(mh,p\tau)}{\partial t}+a_1h\,\frac{\partial u(mh,p\tau)}{\partial x}+O(a^0\tau^2+a_4h^2). \quad (8)$$

Поскольку нашей целью является такой подбор коэффициентов a^0, a_0, a_1, чтобы выполнялось условие аппроксимации (6), то естественно предварительно так сгруппировать слагаемые в правой части равенства (8), чтобы выделился член (7). Тогда остальные слагаемые образуют остаточный член аппроксимации, который должен быть мал. Чтобы выделить член Λu, можно заменить в правой части равенства (8) производные $\partial u/\partial t$ или $\partial u/\partial x$ соответственно по одной из формул

$$\frac{\partial u}{\partial t}\equiv\Lambda u+\frac{\partial u}{\partial x} \quad \text{или} \quad \frac{\partial u}{\partial x}=\frac{\partial u}{\partial t}-\Lambda u.$$

Для определенности воспользуемся первой из них.

Кроме того, подчиним шаги τ, h связи $\tau=rh$, где r — какая-нибудь постоянная. После этого равенство (8) примет следующий вид:

$$\Lambda_h[u]_h\big|_{(mh,p\tau)}=a^0rh\Lambda u\big|_{(mh,p\tau)}+(a^0+a_0+a_1)\,u(mh,p\tau)+$$
$$+(a^0r+a_1)h\,\frac{\partial u(mh,p\tau)}{\partial x}+O(a^0r^2h^2)+O(a,h^2). \quad (9)$$

Среди всех гладких функций $u(x,t)$ можно указать такие, для которых u, $\partial u/\partial x$, $\partial u/\partial t$ в любой заранее заданной фиксированной точке принимают любые независимые друг от друга значения. Следовательно, и значения

$$u, \quad \frac{\partial u}{\partial x}, \quad \Lambda u\equiv\frac{\partial u}{\partial t}-\frac{\partial u}{\partial x}=\varphi(x,t)$$

также можно считать независимыми друг от друга. В силу этого из равенства (9) следует, что для выполнения при любой правой части $\varphi(x,t)$ задачи (4) условия аппроксимации

$$\Lambda_h[u]_h\big|_{(mh,p\tau)}=(\Lambda u)_{(mh,p\tau)}+O(h)$$

необходимо, чтобы выполнялись равенства

$$a^0rh=1+O_1(h), \quad a^0+a_0+a_1=O+O_2(h), \quad a^0r+a_1=O+O_3(h),$$

где $O_1(h)$, $O_2(h)$, $O_3(h)$ — какие-нибудь произвольные величины порядка $O(h)$. Положим $O_1(h) = O_2(h) = O_3(h) = 0$. Получающаяся при этом система

$$a^0 rh = 1, \quad a^0 + a_0 + a_1 = 0, \quad a^0 r + a_1 = 0$$

имеет единственное решение

$$a^0 = \frac{1}{rh} = \frac{1}{\tau}, \quad a_0 = \frac{r-1}{rh} = \frac{1}{h} - \frac{1}{\tau}, \quad a_1 = -\frac{1}{h},$$

которое приводит к уже известной схеме (5).

Теперь мы дополнительно узнали, что среди разностных схем вида

$$L_h u^{(h)} = \begin{cases} a^0 u_m^{p+1} + a_0 u_m^p + a_1 u_{m+1}^p = \varphi(mh, p\tau), \\ u_m^0 = \psi(mh) \end{cases}$$

она является единственной, аппроксимирующей рассматриваемую задачу Коши. Говоря о единственности, мы пренебрегаем тем произволом, который вносит свобода выбора функций $O_1(h)$, $O_2(h)$, $O_3(h)$. Всюду в дальнейших примерах также будем пренебрегать подобного рода очевидным произволом и даже не всегда будем вводить произвольные величины, аналогичные величинам $O_1(h)$, $O_2(h)$, $O_3(h)$, с самого начала полагая их равными нулю.

Посмотрим теперь, как можно строить для задачи (4) разностные схемы

$$L_h u^{(h)} \equiv \begin{cases} a^0 u_m^{p+1} + a_0 u_m^p + a_{-1} u_{m-1}^p + a_1 u_{m+1}^p = \varphi(mh, p\tau), \\ u_m^0 = \psi(mh) \end{cases} \tag{10}$$

более общего вида, связывающие значения искомой функции в четырех точках. Шаги сетки снова свяжем равенством $\tau = rh$ ($r = \text{const}$) и введем обозначение Λ_h, положив

$$\Lambda_h u^{(h)} \equiv a^0 u_m^{p+1} + a_0 u_m^p + a_{-1} u_{m-1}^p + a_1 u_{m+1}^p. \tag{11}$$

Для всякой достаточной гладкой функции $u(x, t)$ с помощью формулы Тейлора можно записать

$$\begin{aligned} \Lambda_h [u]_h = (a^0 + a_0 + a_{-1} + a_1)\, u(mh, p\tau) + a^0 rh \,\frac{\partial u(mh, p\tau)}{\partial t} + \\ + (a_1 - a_{-1}) h \,\frac{\partial u(mh, p\tau)}{\partial x} + \frac{1}{2}\, a^0 r^2 h^2 \,\frac{\partial^2 u(mh, p\tau)}{\partial t^2} + \\ + \frac{1}{2}\, (a_1 + a_{-1}) h^2 \,\frac{\partial^2 u(mh, p\tau)}{\partial x^2} + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3). \end{aligned} \tag{12}$$

Выделим в правой части этого равенства член $\Lambda_u \equiv \partial u / \partial t - \partial u / \partial x$, воспользовавшись для этого тождеством $u_t = u_x + \Lambda u$. Тогда

$$\Lambda_h [u]_h = a^0 rh \Lambda u\big|_{(mh, p\tau)} + (a^0 + a_0 + a_1 + a_{-1})\, u(mh, p\tau) +$$

$$+ (a^0 r + a_1 - a_{-1}) h \frac{\partial u(mh, p\tau)}{\partial x} + \frac{1}{2} a^0 r^2 h^2 \frac{\partial^2 u(mh, p\tau)}{\partial t^2} +$$
$$+ \frac{1}{2} (a_1 + a_{-1}) h^2 \frac{\partial^2 u(mh, p\tau)}{\partial x^2} + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3).$$

Если предполагать, что величина $O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$ достаточно мала (это предположение подтвердится в дальнейшем), то для выполнения условия аппроксимации

$$\Lambda_h [u]_h \big|_{(mh, p\tau)} = \Lambda u \big|_{(mh, p\tau)} + O(h)$$

необходимо, чтобы четыре числа a^0, a_0, a_1, a_{-1} удовлетворяли следующим трем равенствам:

$$\begin{aligned} a^0 r h &= 1, \\ a^0 + a_0 + a_1 + a_{-1} &= 0, \\ a^0 r + a_1 - a_{-1} &= 0. \end{aligned} \tag{13}$$

Система (13) имеет много решений — семейство решений, зависящее от одного параметра. Одно из этих решений

$$a^0 = \frac{1}{rh}, \quad a_{-1} = 0, \quad a_0 = \frac{r-1}{h}, \quad a_1 = -\frac{1}{h}$$

дает уже рассмотренную схему (5). Решению

$$a^0 = \frac{1}{rh}, \quad a_0 = -\frac{1}{rh}, \quad a_{-1} = \frac{1}{2h}, \quad a_1 = -\frac{1}{2h}$$

соответствует схема

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_{m+1}^p - u_{m-1}^p}{2h} = \varphi(mh, p\tau), \\ u_m^0 = \varphi(mh). \end{cases} \tag{14}$$

Выбрав какое-либо решение системы (13), надо его подставить в остаточный член и убедиться, что он мал. Для двух сейчас приведенных решений подстановка чисел a^0, a_0, a_1, a_{-1} дает остаточные члены

$$\frac{a^0 r^2 h^2}{2} \frac{\partial^2 u}{\partial t^2} + \frac{a_1 + a_{-1}}{2} h^2 \frac{\partial^2 u}{\partial x^2} + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$$

порядка $O(h)$.

Среди гладких функций $u(x, t)$ есть многочлены второй степени, для которых $\partial^2 u / \partial t^2$, $\partial^2 u / \partial x^2$ принимают в любой фиксированной точке любые независимые наперед заданные значения. При этом член $O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3)$, в который входят третьи производные многочлена $u(x, t)$, обращается в нуль. Поэтому для того чтобы остаточный член был порядка h, необходимо, чтобы коэффициенты при $\partial^2 u / \partial t^2$,

$\partial^2 u/\partial x^2$ каждый в отдельности были порядка h. Поскольку из первого уравнения (13) имеем $a^0 = 1/(rh)$, то коэффициент при $\partial^2 u/\partial t^2$ есть $rh/2$, и порядок остаточного члена всегда не выше первого.

Мы установили, что нельзя построить разностную схему вида (10), которая аппроксимирует задачу (4) с порядком h^2. Для увеличения порядка аппроксимации пришлось бы увеличить число точек разностной сетки, используемых в конструируемой схеме.

При дополнительном предположении, что $\varphi(x,t) \equiv 0$, существует одна и только одна схема вида (10), аппроксимирующая задачу (4) со вторым порядком относительно h. Построим ее, одновременно убеждаясь в единственности. Заметим, что из тождества $\partial u/\partial t - \partial u/\partial x = 0$ вытекает тождество $\partial^2 u/\partial t^2 = \partial^2 u/\partial x^2$. Действительно,

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right) = \frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial t}\right) = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right).$$

Поэтому при выполнении условий (13), необходимых для аппроксимации первого порядка, равенство (14) примет вид

$$\Lambda_h[u]_h\big|_{(mh,p\tau)} = \frac{1}{2}\,(a^0 r^2 + a_1 + a_{-1}) h^2 u_{xx}\big|_{(mh,p\tau)} + \\ + O(a^0 r^3 h^3, a_1 h^3, a_{-1} h^3). \quad (15)$$

Для получения схемы второго порядка с учетом $a^0 = 1/(rh)$ необходимо, чтобы

$$a^0 r^2 + a_1 + a_{-1} = 0. \quad (16)$$

Система (13), (16) имеет единственное решение

$$a^0 = \frac{1}{rh}, \quad a_0 = \frac{1}{rh} + \frac{r}{h}, \quad a_{-1} = \frac{1-r}{2h}, \quad a_1 = -\frac{1+r}{2h}. \quad (17)$$

Второе слагаемое в правой части (15) есть $O(h^2)$, так что схема (10) с коэффициентами (17) есть единственная схема второго порядка аппроксимации для задачи (4) при $\varphi(x,t) \equiv 0$. Соответствующая разностная схема имеет вид

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{r}{2h}(u_{m+1}^p - 2u_m^p + u_{m-1}^p) = 0. \quad (18)$$

Во всех рассмотренных до сих пор в этой главе примерах разностных схем $L_h u^{(h)} = f^{(h)}$ оператор L_h, который отображает пространство U_h в пространство F_h, задается явными формулами. Но часто оказываются полезными разностные схемы, в которых оператор L_h описывается тем или иным более сложным образом. В дальнейшем мы еще встретимся с задачами, где такие схемы возникают естественным образом.

Изложенные приемы построения разностных схем остаются применимыми и в случае задач с переменными коэффициентами, в случае нелинейных задач, в случае сеток с переменным шагом. Например, в случае неравномерной прямоугольной сетки для замены производных, входящих в дифференциальное уравнение $u_{xx} + u_{yy} = \varphi(x, y)$, разностными отношениями с целью построения разностной схемы можно воспользоваться формулами

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{x_m, y_n} \approx \frac{\dfrac{u(x_{m+1}, y_n) - u(x_m, y_n)}{\Delta x_m} - \dfrac{u(x_m, y_n) - u(x_{m-1}, y_n)}{\Delta x_{m-1}}}{\dfrac{\Delta x_m + \Delta x_{m-1}}{2}},$$

$$\left.\frac{\partial^2 u}{\partial y^2}\right|_{x_m, y_n} \approx \frac{\dfrac{u(x_m, y_{n+1}) - u(x_m, y_n)}{\Delta y_n} - \dfrac{u(x_m, y_n) - u(x_m, y_{n-1})}{\Delta y_{n-1}}}{\dfrac{\Delta y_n + \Delta y_{n-1}}{2}}.$$

Методом неопределенных коэффициентов можно убедиться в единственности этих формул: с точностью до несущественного произвола есть только один набор коэффициентов a_{-1}, a_0, a_1, при котором для любой достаточно гладкой функции $u(x, y)$ имеет место формула

$$\frac{\partial^2 u(x_m, y_n)}{\partial x^2} = a_{-1} u(x_{m-1}, y_n) + a_0 u(x_m, y_n) + \\ + a_1 u(x_{m+1}, y_n) + O(\max(\Delta x_{m-1}, \Delta x_m))$$

с остаточным членом первого порядка малости относительно $\max(\Delta x_{m-1}, \Delta x_m)$. Формулы такого вида с остаточным членом второго порядка малости при $\Delta x_m \neq \Delta x_{m-1}$ не существует.

Для более точной замены производной разностным отношением здесь необходимо привлечь более трех точек сетки.

3. Идея метода А.С. Холодова. В этом методе даются некоторая точная постановка и способы решения следующей задачи. Рассматривается класс схем, имеющих заданный шаблон. Пусть поставлены два требования, которым должна удовлетворять искомая схема, причем одно из них считается основным, а второе — дополнительным. Эти требования могут быть несовместимыми.

Задача состоит в том, чтобы среди схем из заданного класса, удовлетворяющих основному требованию, найти ту (или те), для которой дополнительное требование нарушается меньше, чем для всех других.

Для изложения идеи воспользуемся классом схем вида

$$u_m^{p+1} = \sum_{-j_{\text{лев.}} \leqslant j \leqslant j_{\text{прав.}}} b_j u_{m+j}^p, \tag{19}$$

где $j_{\text{лев.}} \geqslant 0$, $j_{\text{прав.}} \geqslant 0$ — целые числа. Схема вида (10) в случае $\varphi(x, t) \equiv 0$ приводится к виду (19), если положить $j_{\text{лев.}} = 1$, $j_{\text{прав.}} = 1$:

$$u_m^{p+1} = b_{-1} u_{m-1}^p + b_0 u_m^p + b_1 u_{m+1}^p,$$

$$b_{-1} = -\frac{a_{-1}}{a^0}, \quad b_0 = -\frac{a_0}{a^0}, \quad b_1 = -\frac{a_1}{a^0}.$$

Отождествим с каждой схемой вида (19) точку $\mathbf{b} = (b_{-j_{\text{лев.}}}, b_{-j_{\text{лев.}}+1}, \ldots, b_{j_{\text{прав.}}})$ с координатами $\{b_j\}$ в k-мерном линейном пространстве, $k = j_{\text{лев.}} + j_{\text{прав.}} + 1$.

Допустим, что основное требование состоит в том, чтобы схема удовлетворяла некоторому условию $A_{\text{осн.}}$, а дополнительное — в том, чтобы она удовлетворяла также некоторому условию $A_{\text{доп.}}$.

Через M_0 обозначим совокупность точек в пространстве разностных схем, для которых выполнено условие $A_{\text{осн.}}$, а через M_D — совокупность точек, удовлетворяющих условию $A_{\text{доп.}}$.

А.С. Холодов вводит в пространстве разностных схем какую-нибудь норму, после чего за меру невыполнения условия $A_{\text{доп.}}$ для заданной схемы $\mathbf{b} \in M_0$ принимает расстояние $\operatorname{dist}(\mathbf{b}, M_D) = \inf\limits_{x \in M_D} ||\mathbf{b} - \mathbf{x}||$ от точки $\mathbf{b} \in M_0$ до множества M_D. Сформулированная выше задача приобретает тогда точную постановку: среди разностных схем, удовлетворяющих условию $A_{\text{осн.}}$, т. е. среди точек $\mathbf{b}$, принадлежащих множеству M_0, найти ту точку, которая наименее удалена от множества M_D. Эта задача решается средствами линейного программирования, оптимизации и другими.

Выбор нормы остается в руках исследователя и осуществляется с учетом содержательного смысла дополнительного условия и простоты отыскания нужной разностной схемы. Подробное изложение метода и приложения см. в [11].

4. Схемы с пересчетом, или схемы "предиктор-корректор". При построении разностных схем, аппроксимирующих нестационарные задачи, может быть использована та же идея, которая лежит в основе конструкции схем Рунге–Кутта для обыкновенных дифференциальных уравнений,— идея пересчета. Пересчет позволяет повысить порядок аппроксимации, получаемый по исходной схеме, не использующей пересчета. Кроме того, в случае квазилинейных дифференциальных уравнений пересчет дает дополнительную возможность получения так называемых дивергентных схем, о которых будет идти речь в гл. 10.

Напомним идею пересчета на примере простейшей из схем Рунге–Кутта численного решения задачи Коши для обыкновенного дифференциального уравнения

$$\frac{dy}{dt} = f(t, y), \quad y(0) = \psi, \quad 0 < t < T. \tag{20}$$

Если значение y_p в точке $t_p = p\tau$ уже вычислено, то для вычисления y_{p+1} находим предварительно вспомогательную величину $\widetilde{y}_{p+1/2}$, пользуясь простейшей схемой Эйлера (схема "предиктор")

$$\frac{\widetilde{y}_{p+1/2} - y_p}{\tau/2} = f(t_p, \widetilde{y}_p), \tag{21}$$

а затем осуществляем корректирующий пересчет по схеме

$$\frac{y_{p+1}-y_p}{\tau} = f(t_{p+1/2}, \widetilde{y}_{p+1/2}). \qquad (22)$$

Вспомогательная величина $\widetilde{y}_{p+1/2}$, найденная по схеме первого порядка точности, позволяет приближенно найти угловой коэффициент интегральной кривой в середине отрезка $[t_p, t_{p+1}]$ и получить y_{p+1} по формуле (22) с большей точностью, чем это было бы сделано по схеме Эйлера (21).

Мы уже отмечали в § 4 гл. 8, что все соображения остаются в силе, если y, y_p, $\widetilde{y}_{p+1/2}$ будут конечномерными векторами, а f — вектор-функцией. Но можно пойти и дальше, а именно считать y, y_p, $\widetilde{y}_{p+1/2}$ элементами функционального пространства, а f — операторами в этом пространстве.

Например, задачу Коши

$$\begin{gathered}\frac{\partial u}{\partial t} + A\,\frac{\partial u}{\partial x} = 0, \quad -\infty < x < \infty, \quad 0 < t < T, \\ u(x,0) = \psi(x), \quad -\infty < x < \infty,\end{gathered} \qquad (23)$$

$A = \mathrm{const}$, можно считать задачей вида (20), если положить $y(t) = u(x,t)$, так что при каждом t под y надо понимать функцию аргумента x, а под операцией f — оператор $-A\partial/\partial x$. Приведем пример разностной схемы с пересчетом для задачи (23).

П р и м е р. Пусть сеточная функция $u^p = \{u_m^p\}$ $(m = 0, \pm 1, \ldots)$ при данном p уже вычислена. Найдем вспомогательную сеточную функцию $\widetilde{u}^{p+1/2} = \{u_m^{p+1/2}\}$ $(m = 0, \pm 1, \ldots)$, отнесенную к моменту времени $t_{p+1/2} = (p+1/2)\tau$ и к точкам $x_{m+1/2} = (m+1/2)h$, воспользовавшись следующей схемой первого порядка точности:

$$\frac{\widetilde{u}_{m+1/2}^{p+1/2} - (u_{m+1}^p + u_m^p)/2}{\tau/2} + A\,\frac{u_{m+1}^p - u_m^p}{h} = 0, \quad m = 0, \pm 1, \ldots \qquad (24)$$

Затем осуществим коррекцию и найдем u^{p+1} с помощью схемы

$$\frac{u_m^{p+1} - u_m^p}{\tau} + A\,\frac{\widetilde{u}_{m+1/2}^{p+1/2} - \widetilde{u}_{m-1/2}^{p+1/2}}{h} = 0, \quad m = 0, \pm 1, \ldots \qquad (25)$$

Исключая $\widetilde{u}^{p+1/2}$ из уравнений (24), (25), получаем схему

$$\begin{gathered}\frac{u_m^{p+1} - u_m^p}{\tau} + A\,\frac{u_{m+1}^p - u_{m-1}^p}{2h} + A^2\frac{\tau}{2}\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ u_m^0 = \psi(mh), \quad m = 0, \pm 1, \ldots; \quad p = 0, 1, \ldots, [T/\tau] - 1.\end{gathered} \qquad (26)$$

Эта схема при $A = -1$ совпадает со схемой (18). Случай $A \neq 1$ несущественно отличается от разобранного. Схема (26), а значит, и схема

с пересчетом (24), (25) имеют второй порядок аппроксимации по h; $\tau = rh$ ($r = \text{const}$).

Задачи

1. Для задачи Коши

$$\frac{\partial u}{\partial t} - \left(\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y}\right) = \varphi(x, y, t), \quad -\infty < x, y < \infty, \quad 0 \leqslant t < T,$$
$$u(x, y, 0) = \varphi(x, y), \quad -\infty < x, y < \infty,$$

воспользоваться сеткой $x_m = mh$, $y_n = nh$, $t_p = p\tau$ и построить какую-либо аппроксимирующую ее разностную схему.

2. Для задачи о теплопроводности

$$\begin{gathered} \partial u/\partial t = \partial^2 u/\partial x^2, \quad -\infty < x, y < \infty, \quad 0 \leqslant t \leqslant T, \\ u(x, 0) = \psi(x), \quad -\infty < x < \infty, \end{gathered} \tag{27}$$

рассмотреть разностную схему

$$\frac{u_m^{p+1} - u_m^p}{\tau} = \sigma\,\frac{u_{m+1}^{p+1} - 2u_m^{p+1} + u_{m-1}^{p+1}}{h^2} + (1 - \sigma)\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2},$$
$$u_m^0 = \psi(mh),$$

где σ — параметр.

а) Показать, что при любом σ имеет место аппроксимация на гладком решении $u(x, t)$ с порядком $O(\tau + h^2)$.

б) Подобрать σ так, чтобы аппроксимация была $O(\tau^2 + h^2)$.

в) Связав шаги сетки соотношением $\tau/h^2 = r = \text{const}$, подобрать затем σ так, чтобы получить аппроксимацию порядка h^4.

г) При $\sigma = 0$ подобрать число $r = \tau/h^2$ так, чтобы аппроксимация имела порядок h^4.

д*) Можно ли за счет выбора σ при фиксированном $\tau/h^2 = r$ добиться того, чтобы аппроксимация на любом гладком решении была порядка выше четвертого?

3. Для задачи о теплопроводности

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(x, t)\,\frac{\partial u}{\partial x}\right], \quad -\infty < x < \infty, \quad 0 < t < T,$$
$$u(x, 0) = \psi(x), \quad -\infty < x < \infty,$$

пользуясь сеткой $x_m = mh$, $t_p = p\tau$, построить аппроксимирующую ее разностную схему.

4. Для нелинейной задачи о теплопроводности

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(u)\,\frac{\partial u}{\partial x}\right], \quad -\infty < x < \infty, \quad 0 \leqslant t \leqslant T,$$
$$u(x, 0) = \psi(x),$$

пользуясь сеткой $x_m = mh$, $t_p = p\tau$, построить аппроксимирующую ее явную разностную схему. Выписать формулы для вычисления $u^{(h)}$ по этой схеме.

5. Доказать, что схема с пересчетом для задачи (27), в которой значения решения $\widetilde{u}_m^{p+1/2}$ на промежуточном слое определяются по неявной схеме порядка аппроксимации $O(\tau + h^2)$

$$\frac{\widetilde{u}_m^{p+1/2} - u_m^p}{\tau/2} - \frac{\widetilde{u}_{m+1}^{p+1/2} - 2\widetilde{u}_m^{p+1/2} + \widetilde{u}_{m-1}^{p+1/2}}{h^2} = 0, \quad m = 0, \pm 1, ...,$$

а решение $\{u_m^{p+1}\}$ — по схеме

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{\widetilde{u}_{m+1}^{p+1/2} - 2\widetilde{u}_m^{p+1/2} + \widetilde{u}_{m-1}^{p+1/2}}{h^2} = 0,$$
$$u_m^0 = \psi(mh),$$

обладает аппроксимацией порядка $O(\tau^2 + h^2)$ на гладком решении u.

6. Изменим разностную схему (10), отвечающую коэффициентам (17), заменив в правой части $\varphi(mh, p\tau)$ на

$$\varphi_m^p = \varphi(mh, p\tau) + \frac{rh}{2}(\varphi_t + \varphi_x)\Big|_{\substack{x=mh\\ t=p\tau}}.$$

Показать, что полученная при этом схема аппроксимирует задачу (4) с порядком h^2 при произвольной правой части $\varphi(x,t)$, имеющей ограниченные производные достаточно высокого порядка.

7. Записать схему (10) при $\varphi(x,t)$ в виде

$$u_m^{p+1} = b_{-1}u_{m-1}^p + b_0 u_m^p + b_1 u_{m+1}^p. \qquad (10')$$

Исходную схему (10) будем называть *монотонной*, если $b_j \geqslant 0$ $(j = -1, 0, 1)$.

Будем считать, что условие $A_0 = A_{\text{осн.}}$ состоит в том, чтобы схема (10) имела порядок аппроксимации не ниже первого и была монотонна.

Описать в трехмерном пространстве R множество точек $(b_{-1},\ b_0,\ b_1)$, которому соответствуют схемы, удовлетворяющие условию A_0.

Ответ. В случае $r = \tau/h > 1$ множество M_0 пусто. В случае $r = 1$ оно состоит из одной точки $(0,0,1)$. В случае $0 < r < 2$ множество M_0 есть отрезок с концами $\left(\frac{1-r}{2}, 0, \frac{1+r}{2}\right)$, $(0, 1-r, r)$.

8. Для задачи Коши для уравнения колебаний струны

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = \varphi(x,t), \quad -\infty < x < \infty, \quad 0 \leqslant t \leqslant T,$$
$$u(x,0) = \psi^{(0)}(x), \quad \frac{\partial u(x,0)}{\partial t} = \psi^{(1)}(x), \quad -\infty < x < \infty,$$

исследовать аппроксимацию, которой обладает на достаточно гладком решении $u(x,t)$ разностная схема

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{p+1} - 2u_m^p + u_m^{p-1}}{\tau^2} - \dfrac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \varphi(mh, p\tau), \\ u_m^{(0)} = \psi^{(0)}(mh), \\ \dfrac{u_m^1 - u_m^0}{\tau} = \psi^{(1)}(mh). \end{cases}$$

За норму $\|f^{(h)}\|_{F_h}$ принять максимум модулей всех компонент элемента

$$f^{(h)} = \begin{cases} \varphi_m^p, \\ \psi_m^{(0)}, \\ \psi_m^{(1)}. \end{cases}$$

Показать, что аппроксимация имеет первый порядок относительно h, если $\tau = rh$ ($r = \text{const}$).

Как следует задать значения $\psi_m^{(1)}$, используя заданные функции $\varphi(x, y, t)$, $\psi^{(0)}(x)$, $\psi^{(1)}(x)$, чтобы порядок аппроксимации оказался вторым?

9. Для задачи о распространении тепла на отрезке

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \varphi(x, t), \quad 0 \leqslant x \leqslant 1, \quad 0 \leqslant t \leqslant T,$$
$$u(x, 0) = \psi^{(0)}(x), \quad 0 \leqslant x \leqslant 1,$$
$$u(0, t) = \psi^{(1)}(t), \quad 0 \leqslant t \leqslant 1,$$
$$u(1, 1) = \psi^{(2)}(t), \quad 0 \leqslant t \leqslant 1,$$

рассмотреть разностную схему вида

$$\begin{gathered} \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \varphi(mh, p\tau), \\ m = 1, 2, ..., M-1; \quad h = M^{-1}, \quad p = 0, 1, ..., [T/\tau] - 1, \\ u_m^0 = \psi^{(0)}(mh), \quad m = 0, 1, ..., M, \\ u_0^p = \psi^{(1)}(p\tau), \quad p = 0, 1, ..., [T/\tau], \\ u_M^p = \psi^{(2)}(p\tau), \quad p = 0, 1, ..., [T/\tau]. \end{gathered} \tag{11}$$

За норму $\|\cdot\|_{F_h}$ принять максимум абсолютных величин правых частей уравнений, составляющих в совокупности рассматриваемую разностную схему. Шаги τ, h считать связанными равенством $\tau = rh^2$.

Показать, что схема (11) обладает порядком аппроксимации h^2.

§ 3. Спектральный признак устойчивости разностной задачи Коши

Изложим широко применяемый способ Неймана исследования разностных задач с начальными данными. В этом параграфе ограничимся случаем разностной задачи Коши с постоянными коэффициентами, а в § 4 частично распространим результаты на случай переменных коэффициентов и смешанных задач.

1. Устойчивость по начальным данным. Простейшим примером разностной задачи Коши может служить неоднократно встречавшаяся выше задача

$$L_h u^{(h)} \equiv \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_{m+1}^p - u_m^p}{h} = \varphi_m^p, \quad p = 0, 1, ..., \left[\dfrac{T}{\tau}\right] - 1, \\ u_m^{(0)} = \psi_m, \quad m = 0, \pm 1, ... \end{cases} \tag{1}$$

Положив

$$f^{(h)} = \begin{cases} \varphi_m^p, & p = 0, 1, ..., [T/\tau] - 1, \\ \psi_m, & m = 0, \pm 1, ..., \end{cases}$$

запишем задачу (1) в форме

$$L_h u^{(h)} = f^{(h)}. \tag{2}$$

Определим нормы $\|u^{(h)}\|_{U_h}$ и $\|f^{(h)}\|_{F_h}$ равенствами

$$\|u^{(h)}\|_{U_h} = \sup_{m,p} |u_m^p|, \quad \|f^{(h)}\|_{F_n} = \sup_{m,p} |\varphi_m^p| + \sup_m |\psi_m|.$$

Тогда условие устойчивости задачи (2)

$$\|u^{(h)}\|_{U_h} \leqslant c \, \|f^{(h)}\|_{F_h} \tag{3}$$

примет вид

$$\max_m |u_m^p| \leqslant c \Big[\max_m |\psi_m| + \max_{m,k} |\varphi_m^k| \Big], \quad p = 0, 1, ..., [T/\tau], \tag{4}$$

где c не зависит от h (и от $\tau = rh$, $r = \text{const}$). Условие (4) должно выполняться при произвольных $\{\psi_m\}$, $\{\varphi_m^p\}$. В частности, для устойчивости необходимо, чтобы оно выполнялось при произвольных $\{\psi_m\}$ и $\varphi_m^p \equiv 0$, т. е. чтобы решение задачи

$$\begin{gathered} \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad p = 0, 1, ..., \left[\frac{T}{\tau}\right] - 1, \\ u_m^0 = \psi_m, \quad m = 0, \pm 1, ..., \end{gathered} \tag{5}$$

удовлетворяло условию

$$\max_m |u_m^p| \leqslant c \max_m |\psi_m|, \quad p = 0, 1, ..., [T/\tau], \tag{6}$$

при произвольной ограниченной функции ψ_m

Свойство (6), необходимое для устойчивости (4) задачи (1), называют *устойчивостью задачи* (1) *относительно возмущения начальных данных.* Оно означает, что возмущение ψ_m, внесенное в начальные данные задачи (1), вызовет возмущение решения задачи (1), которое в силу (6) не более, чем в c раз превосходит возмущение начальных данных, причем c не зависит от h.

2. Необходимое спектральное условие устойчивости. Для устойчивости задачи Коши (1) по начальным данным необходимо, чтобы условие (6) выполнялось, в частности, если ψ_m есть какая-нибудь гармоника

$$u_m^0 = \psi_m = e^{i\alpha m}, \quad m = 0, \pm 1, ..., \tag{7}$$

где α — вещественный параметр. Но решение задачи (5) при начальном условии (7) имеет вид

$$u_m^p = \lambda^p e^{i\alpha m}, \tag{8}$$

где $\lambda = \lambda(\alpha)$ определяется путем подстановки выражения (8) в однородное разностное уравнение задачи (5):

$$\lambda(\alpha) = 1 - r + re^{i\alpha}, \quad r = \tau/h = \text{const}. \tag{9}$$

Для решения (8) справедливо равенство

$$\max_m |u_m^p| = |\lambda(\alpha)|^p \max_m |u_m^0| = |\lambda(\alpha)|^p \max_m |\psi_m|.$$

Поэтому для выполнения условия (6) необходимо, чтобы при всех вещественных α выполнялось неравенство

$$|\lambda(\alpha)|^p \leqslant c, \quad p = 0, 1, ..., [T/\tau],$$

или

$$|\lambda(\alpha)| \leqslant 1 + c_1\tau, \tag{10}$$

где c_1 — некоторая постоянная, не зависящая от α, τ. Это и есть необходимое спектральное условие Неймана применительно к рассматриваемому примеру. Спектральным оно называется по следующей причине.

Существование решения вида (8) показывает, что гармоника $\{e^{i\alpha m}\}$ является собственной функцией оператора перехода

$$u_m^{p+1} = (1 - r)u_m^p + ru_{m+1}^p, \quad m = 0, \pm 1, ...,$$

который в силу разностного уравнения (5) ставит в соответствие сеточной функции $\{u_m^p\}$ $(m = 0, \pm 1, ...)$, определенной на слое $t = t_p$, сеточную функцию $\{u_m^{p+1}\}$, определенную на слое $t = t_{p+1}$. Число $\lambda(\alpha) = 1 - r + re^{i\alpha}$ является соответствующим этой гармонике $\{e^{i\alpha m}\}$ собственным числом оператора перехода. Линия, которую пробегает точка $\lambda = \lambda(\alpha)$ на комплексной плоскости, когда α пробегает вещественную ось, вся состоит из собственных значений и является спектром оператора перехода.

Таким образом, необходимое условие устойчивости (10) можно сформулировать так: спектр оператора перехода, соответствующего разностному уравнению задачи (5), должен лежать в круге радиуса $1 + c_1\tau$ на комплексной плоскости. В нашем примере спектр (9) не зависит от τ. Поэтому условие (10) равносильно требованию, чтобы спектр $\lambda(\alpha)$ лежал в единичном круге:

$$|\lambda(\alpha)| \leqslant 1. \tag{11}$$

Воспользуемся сформулированным признаком для анализа устойчивости задачи (1). Спектр (9) представляет собой окружность с центром в точке $1 - r$ и радиусом r на комплексной плоскости. В случае

$r < 1$ эта окружность лежит внутри единичного круга (касаясь его в точке $\lambda = 1$), при $r = 1$ совпадает с единичной окружностью, а при

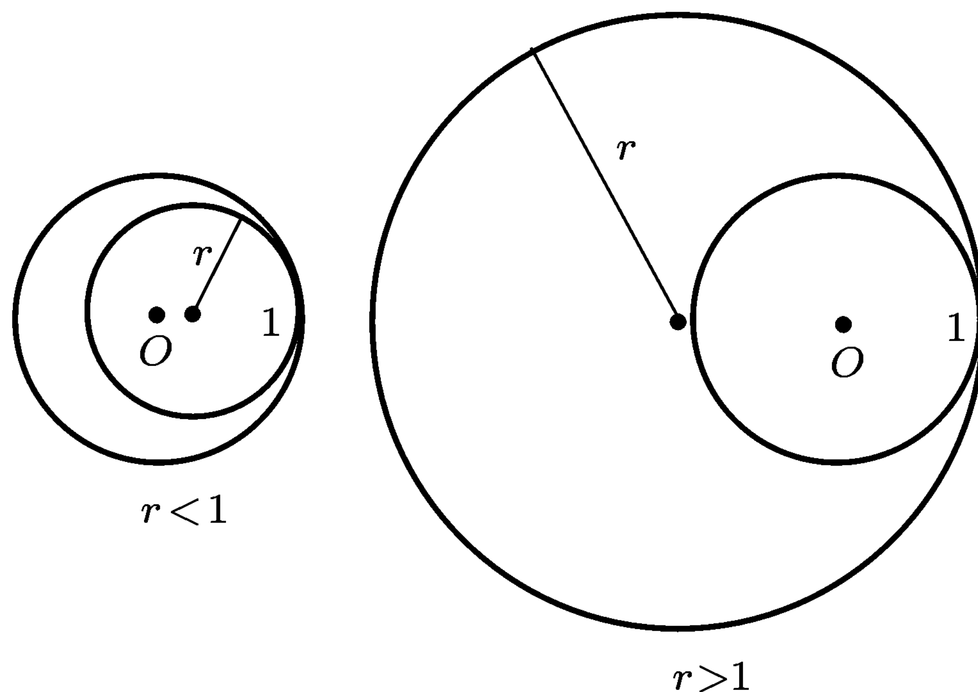


Рис. 20

$r > 1$ лежит вне единичного круга (рис. 20). Соответственно необходимое условие устойчивости (11) выполнено при $r \leqslant 1$ и не выполнено при $r > 1$. В п. 3 § 1 мы исследовали рассматриваемую разностную схему и показали, что при $r \leqslant 1$ она устойчива, а при $r > 1$ неустойчива. Таким образом, необходимое условие устойчивости Неймана оказалось в данном случае достаточно чувствительным, чтобы в точности отделить случай устойчивости от случая неустойчивости.

В общем случае задачи Коши для разностных уравнений и систем разностных уравнений необходимый спектральный признак устойчивости Неймана состоит в том, что спектр $\lambda = \lambda(\alpha, h)$ разностной задачи при всех достаточно малых h должен лежать в круге

$$|\lambda| \leqslant 1 + \varepsilon \tag{12}$$

на комплексной плоскости, как бы мало ни было заранее выбранное положительное число ε.

Заметим, что если для рассматриваемой разностной задачи спектр окажется не зависящим от h (и от τ), то условие (12) равносильно требованию, чтобы спектр $\lambda = \lambda(\alpha, h) = \lambda(\alpha)$ лежал в единичном круге (11).

Под спектром разностной задачи в условии (12) понимается совокупность всех $\lambda = \lambda(\alpha, h)$, при которых соответствующее однородное разностное уравнение (или система уравнений) имеет решение вида

$$u_m^p = \lambda^p [u^0 e^{i\alpha m}], \tag{13}$$

где u^0 — число (единица), если речь идет о скалярном разностном уравнении, и числовой вектор, если речь идет о векторном разностном уравнении, т. е. о системе скалярных разностных уравнений.

Если необходимое условие Неймана (12) не выполнено, то ни при каком разумном выборе норм нельзя ожидать устойчивости, а в случае его выполнения можно надеяться, что при некотором разумном выборе норм устойчивость имеет место.

3. Примеры. Рассмотрим ряд интересных разностных задач Коши и применим для анализа их устойчивости спектральный признак Неймана. Начнем с разностных схем, аппроксимирующих дифференциальную задачу Коши

$$\begin{gathered}\frac{\partial u}{\partial t}-\frac{\partial u}{\partial x}=\varphi(x,t), \quad -\infty<x<\infty, \quad 0<t<T,\\ u(x,0)=\psi(x), \quad -\infty<x<\infty.\end{gathered} \tag{14}$$

Пример 1. Рассмотрим разностную схему

$$\begin{gathered}\frac{u_m^{p+1}-u_m^p}{\tau}-\frac{u_m^p-u_{m-1}^p}{h}=\varphi(mh,p\tau), \quad m=0,\pm1,\dots,\\ u_m^0=\psi(mh), \quad p=0,1,\dots,[T/\tau]-1.\end{gathered}$$

Подставляя выражение вида (8) в соответствующее однородное разностное уравнение, после простых преобразований получаем

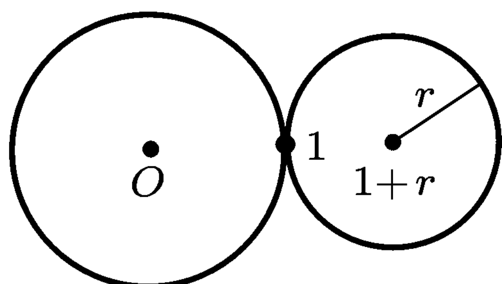


Рис. 21

$$\lambda(\alpha)=1+r-re^{-i\alpha}.$$

В силу этой формулы спектр представляет собой окружность с центром в точке $1+r$ и радиусом r (рис. 21). Ни при каком r спектр не лежит в единичном круге. Условие устойчивости (12) всегда не выполнено. Этого можно было ожидать.

При любом $r=\tau/h$ не выполнено необходимое условие сходимости (и устойчивости) Куранта, Фридрихса и Леви.

Пример 2. Рассмотрим разностную схему

$$\begin{gathered}\frac{u_m^{p+1}-u_m^p}{\tau}-\frac{u_{m+1}^p-u_{m-1}^p}{2h}-\frac{\tau}{2h^2}\left(u_{m-1}^p-2u_m^p+u_{m-1}^p\right)=\\ =\varphi(mh,p\tau), \\ u_m^0=\psi(mh),\end{gathered} \tag{15}$$

аппроксимирующую (проверьте) задачу (14) при $\varphi\equiv0$ со вторым порядком относительно h. Для нее $\lambda=\lambda(\alpha)$ определяется из уравнения

$$\frac{\lambda-1}{\tau}-\frac{e^{i\alpha}-e^{-i\alpha}}{2h}-\frac{\tau}{2h^2}\left(e^{i\alpha}-2+e^{-i\alpha}\right)=0.$$

Обозначим по-прежнему $r=\tau/h$. Заметив, что

$$\frac{e^{i\alpha}-e^{-i\alpha}}{2i}=\sin\alpha,$$

$$\frac{e^{i\alpha}-2+e^{-i\alpha}}{4}=-\left(\frac{e^{i\alpha/2}-e^{-i\alpha/2}}{2i}\right)^2=-\sin^2\frac{\alpha}{2},$$

получим

$$\lambda(\alpha) = 1 + ir\sin\alpha - 2r^2\sin^2\frac{\alpha}{2}, \tag{16}$$

$$|\lambda(\alpha)|^2 = \left(1 - 2r^2\sin^2\frac{\alpha}{2}\right)^2 + r^2\sin^2\alpha.$$

После простых преобразований найдем

$$1 - |\lambda|^2 = 4r^2\sin^4\frac{\alpha}{2}(1 - r^2). \tag{17}$$

Условие Неймана выполнено, если правая часть неотрицательна ($r \leqslant 1$), и не выполнено при $r > 1$.

Пример 3. Рассмотрим разностную схему

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - \dfrac{u_{m+1}^p - u_{m-1}^p}{2h} = \varphi(mh, p\tau), \\ u_m^0 = \psi(mh) \end{cases} \tag{18}$$

для той же задачи Коши (14). Подставляя в уравнение (18) выражение (8), после сокращений получаем уравнение для λ:

$$\frac{\lambda - 1}{\tau} - \frac{e^{i\alpha} - e^{-i\alpha}}{2h} = 0,$$

или

$$\lambda(\alpha) = 1 + i\left(\frac{\tau}{h}\sin\alpha\right).$$

Спектр $\lambda = \lambda(\alpha)$ заполняет вертикальный отрезок длины $2\tau/h$, проходящий через точку $\lambda = 1$ (рис. 22).

Если $\tau/h = r = \text{const}$, то условие (12) не выполняется, спектр не лежит в нужном круге. Если при $h \to 0$ шаг τ изменяется как h^2, так что $r = \tau/h^2$, то самая далекая от точки $\lambda = 0$ точка $\lambda(\alpha)$ имеет модуль

$$|\lambda(\alpha)|_{\alpha=\pi/2} = \sqrt{1 + \left(\frac{\tau}{h}\right)^2} = \sqrt{1 + \tau r} < 1 + \frac{r}{2}\tau.$$

Условие $|\lambda(\alpha)| \leqslant 1 + c\tau$ в этом случае выполнено при $c = r/2$.

Ясно, что требование $\tau = rh^2$ является гораздо более жестким условием на убывание шага по времени τ при стремлении шага h к нулю, чем требование $\tau = rh$, которого было достаточно для выполнения признака Неймана для разностных схем (5), (15), аппроксимирующих ту же задачу Коши (14).

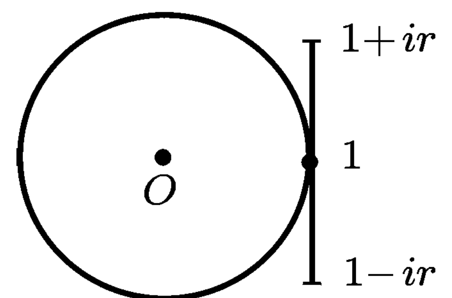


Рис. 22

Отметим, что признак Куранта, Фридрихса и Леви позволяет утверждать (проверьте), что схема неустойчива только при $\tau/h = r > 1$, а при $\tau/h = r \leqslant 1$ суждений об устойчивости не дает и оказывается слабее признака Неймана.

Рассмотрим теперь две разностные схемы, аппроксимирующие задачу Коши для уравнения теплопроводности

$$\begin{gathered} \frac{\partial u}{\partial t} - a^2 \frac{\partial^2 u}{\partial x^2} = \varphi(x,t), \quad -\infty < x < \infty, \quad 0 \leqslant t \leqslant T, \\ u(x,0) = \psi(x), \quad -\infty < x < \infty. \end{gathered} \tag{19}$$

Пример 4. Явная разностная схема

$$\frac{u_m^{p+1} - u_m^p}{\tau} - a^2 \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \varphi(mh, p\tau),$$

$$u_m^0 = \psi(mh), \quad m = 0, \pm 1, ..., \quad p = 0, 1, ..., [T/\tau] - 1,$$

позволяющая вычислить $\{u_m^{p+1}\}$ по явной форме через $\{u_m^p\}$:

$$u_m^{p+1} = (1 - 2r)u_m^p + r(u_{m+1}^p + u_{m-1}^p) + \tau\varphi_m^p,$$

при подстановке $u^p = \lambda^p e^{i\alpha m}$ в соответствующее однородное разностное уравнение приводит к соотношению

$$\frac{\lambda - 1}{\tau} - a^2 \frac{e^{-i\alpha} - 2 + e^{i\alpha}}{h^2} = 0.$$

Заметив, что

$$\frac{e^{i\alpha} - 2 + e^{-i\alpha}}{4} = -\sin^2 \frac{\alpha}{2},$$

получим

$$\lambda(\alpha) = 1 - 4ra^2 \sin^2 \frac{\alpha}{2}, \quad r = \frac{\tau}{h^2}.$$

При изменении α число $\lambda(\alpha)$ пробегает отрезок $1 - 4ra^2 \leqslant \lambda \leqslant 1$ вещественной оси (рис. 23).

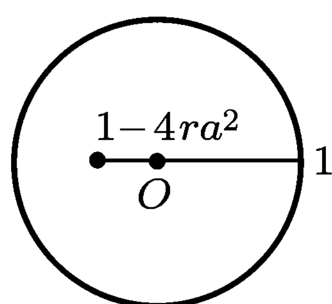


Рис. 23

Для устойчивости необходимо, чтобы левый конец этого отрезка лежал в единичном круге $1 - 4ra^2 \geqslant -1$, или

$$r \leqslant 1/(2a^2). \tag{20}$$

В случае если $r > 1/(2a^2)$, точка

$$\lambda(\alpha) = 1 - 4ra^2 \sin^2(\alpha/2),$$

отвечающая $\alpha = \pi$, лежит левее точки -1. Гармоника $e^{i\pi m} = (-1)^m$ порождает решение

$$u_m^p = (1 - 4ra^2)^p (-1)^m,$$

не удовлетворяющее условию (6) ни при какой постоянной c.

Пример 5. Рассмотрим теперь вторую схему — неявную:

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau} - a^2 \dfrac{u_{m-1}^{p+1} - 2u_m^{p+1} + u_{m+1}^{p+1}}{h^2} = \varphi(mh, p\tau), \\ u_m^0 = \psi(mh), \\ m = 0, \pm 1, ...; \quad p = 0, 1, ..., [T/\tau] - 1. \end{cases} \tag{21}$$

Здесь $\{u_m^{p+1}\}$ не выражается через $\{u_m^p\}$ явной формулой, так как в уравнение при данном m входит не одно, а три неизвестных: u_m^{p+1}, u_m^{p+1}, u_{m+1}^{p+1}. Поэтому схему (21) называют *неявной.*

Аналогичные выкладки приводят к выражению

$$\lambda(\alpha) = \frac{1}{1 + 4ra^2 \sin^2(\alpha/2)}, \quad r = \frac{\tau}{h^2}. \tag{22}$$

Спектр этой задачи заполняет отрезок

$$\left(1 + 4ra^2 \sin^2 \frac{\alpha}{2}\right)^{-1} \leqslant \lambda \leqslant 1$$

вещественной оси, и условие $|\lambda| < 1$ выполнено при любом r.

Спектральный признак Неймана применим для исследования разностной задачи Коши и в случае, если пространственных переменных две или более.

Пример 6. Для задачи

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad -\infty < x, y < \infty, \quad 0 < t < T,$$
$$u(x, y, 0) = \psi(x, y)$$

возьмем сетку $(x_m, y_n, t_p) = (mh, nh, p\tau)$. Заменяя производные разностными отношениями, построим разностную схему

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{mn}^{p+1} - u_{mn}^p}{\tau} - \dfrac{u_{m+1,n}^p - 2u_{mn}^p + u_{m-1,n}^p}{h^2} + \\ \qquad\qquad + \dfrac{u_{m,n+1}^p - 2u_{mn}^p + u_{m,n-1}^p}{h^2} = 0, \\ u_{mn}^0 = \psi(mh, nh), \\ m, n = 0, \pm 1, ...; \quad p = 0, 1, ..., [T/\tau] - 1. \end{cases} \tag{23}$$

Задавая u_{mn}^0 в виде $e^{i(\alpha m + \beta n)}$, т. е. в виде двумерной гармоники, зависящей от двух вещественных параметров α, β, находим решение вида

$$u_{mn}^p = \lambda^p(\alpha, \beta) e^{i(\alpha m + \beta n)}.$$

Подставляя это выражение в разностное уравнение, после сокращений и тождественных преобразований находим

$$\lambda(\alpha, \beta) = 1 - 4r\left(\sin^2 \frac{\alpha}{2} + \sin \frac{\beta}{2}\right).$$

При изменении вещественных α, β точка $\lambda = \lambda(\alpha, \beta)$ пробежит отрезок $1 - 8r \leqslant \lambda \leqslant 1$ вещественной оси. Условие устойчивости выполняется, если $1 - 8r \geqslant -1$, или $r \leqslant 1/4$.

Приведем пример, иллюстрирующий применение признака Неймана для разностных уравнений, связывающих значения искомой функции не на двух, а на трех временных слоях.

Пример 7. Задачу Коши для волнового уравнения

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \quad -\infty < x < \infty, \quad 0 \leqslant t \leqslant T,$$

$$u(x, 0) = \psi^{(0)}(x), \quad \frac{\partial u(x, 0)}{\partial t} = \psi^{(1)}(x), \quad -\infty < x < \infty,$$

аппроксимируем разностной схемой

$$\begin{aligned} &\frac{u_m^{p+1} - 2u_m^p + u_{m-1}^{p-1}}{\tau^2} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ &u_m^0 = \psi^{(0)}(mh), \quad \frac{u_m^1 - u_m^0}{\tau} = \psi^{(1)}(mh), \\ &m = 0, \pm 1, ...; \quad p = 0, 1, ..., [T/\tau] - 1. \end{aligned} \tag{24}$$

Подставляя в разностное уравнение решение вида (8), получаем после простых преобразований следующее уравнение для определения λ:

$$\lambda^2 - 2\Big(1 - 2r^2 \sin^2 \frac{\alpha}{2}\Big)\lambda + 1 = 0, \quad r = \frac{\tau}{h}.$$

Произведение корней этого уравнения равно единице. Если дискриминант

$$d(\alpha) = 4r^2 \sin^2 \frac{\alpha}{2}\Big(r^2 \sin^2 \frac{\alpha}{2} - 1\Big)$$

квадратного уравнения отрицателен, то корни $\lambda_1(\alpha)$, $\lambda_2(\alpha)$ комплексно сопряженные и равные единице по модулю. В случае $r < 1$ дискриминант остается отрицательным при всех α. На рис. 24, *а* изображен

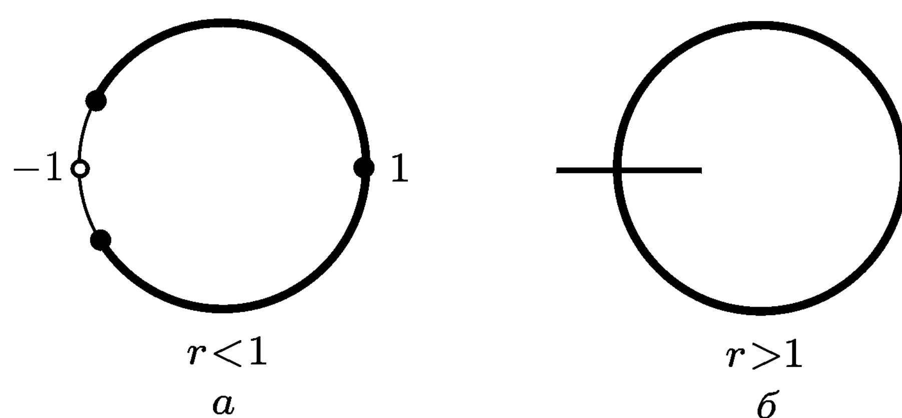


Рис. 24

спектр для этого случая. Он заполняет часть единичной окружности. В случае $r = 1$ спектр заполняет всю окружность. При $r > 1$ по мере увеличения от нуля до π корни $\lambda_1(\alpha)$, $\lambda_2(\alpha)$ движутся из точки $\lambda = 1$ по единичной окружности: один по часовой стрелке, а другой против часовой стрелки, пока не сольются в точке $\lambda = -1$, а затем один из корней пойдет по вещественной оси из точки $\lambda = -1$ влево, а другой вправо, так как они вещественны и $\lambda_1\lambda_2 = 1$ (рис. 24, *б*). Условие устойчивости выполнено при $r \leqslant 1$.

Рассмотрим задачу Коши для следующей гиперболической системы дифференциальных уравнений, описывающей распространение звука:

$$\begin{gathered} \partial v/\partial t = \partial w/\partial x, \quad \partial w/\partial t = \partial v/\partial x, \\ -\infty < x < \infty, \quad 0 \leqslant t \leqslant T, \\ v(x,0) = \psi^{(1)}(x), \quad w(x,0) = \psi^{(2)}(x). \end{gathered} \tag{25}$$

Положим

$$u(x,t) = \begin{bmatrix} v(x,t) \\ w(x,t) \end{bmatrix}, \quad \psi(x) = \begin{bmatrix} \psi^{(1)}(x) \\ \psi^{(2)}(x) \end{bmatrix}$$

и запишем (25) в векторной форме:

$$\begin{gathered} \frac{\partial u}{\partial t} - A\frac{\partial u}{\partial x} = 0, \quad -\infty < x < \infty, \quad 0 \leqslant t \leqslant T, \\ u(x,0) = \psi(x), \quad -\infty < x < \infty, \end{gathered} \tag{25$'$}$$

где

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Исследуем две разностные схемы, аппроксимирующие задачу (25′).

Пример 8. Рассмотрим разностную схему

$$\frac{u_m^{p+1} - u_m^p}{\tau} - A\,\frac{u_{m+1}^p - u_m^p}{h} = 0, \quad u_m^0 = \psi(mh). \tag{26}$$

Ищем решение векторного однородного разностного уравнения в виде (13):

$$u_m^p = \lambda^p \begin{bmatrix} v^0 \\ w^0 \end{bmatrix} e^{i\alpha m}.$$

Подставляя это выражение в разностное уравнение (26), приходим к равенству

$$\frac{\lambda - 1}{\tau}\,u^0 - A\,\frac{e^{i\alpha} - 1}{h}\,u^0 = 0,$$

или

$$(\lambda - 1)u^0 - r(e^{i\alpha} - 1)Au^0, \quad r = \tau/h, \tag{27}$$

которое можно рассматривать как векторную запись системы линейных уравнений для определения компонент вектора u^0.

Запишем систему (27) в развернутой форме:

$$\begin{bmatrix} \lambda - 1 & -r(e^{i\alpha} - 1) \\ -r(e^{i\alpha} - 1) & \lambda - 1 \end{bmatrix} \begin{bmatrix} v^0 \\ w^0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{28}$$

Эта система линейных уравнений имеет нетривиальное решение $u^0 = \begin{bmatrix} v^0 \\ w^0 \end{bmatrix}$ лишь при тех $\lambda = \lambda(\alpha)$, при которых определитель системы (28) обращается в нуль:

$$(\lambda - 1)^2 = r^2(e^{i\alpha} - 1)^2.$$

Отсюда

$$\lambda_1(\alpha) = 1 - r + re^{i\alpha}, \quad \lambda_2(\alpha) = 1 + r - re^{i\alpha}. \tag{29}$$

Корни $\lambda_1(\alpha)$, $\lambda_2(\alpha)$ пробегают окружности радиуса r с центрами в точках $1 - r$, $1 + r$ соответственно (рис. 25). Условие устойчивости Неймана не выполнено ни при каком r.

Пример 9. Рассмотрим разностную схему

$$\frac{u_m^{p+1} - u_m^p}{\tau} - A\,\frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{\tau}{2h^2}\,A^2(u_{m+1}^p - 2u_m^p + u_{m-1}^p) = 0,$$
$$u_m^0 = \psi(mh), \quad m = 0, \pm 1, ..., \quad p = 0, 1, ..., [T/\tau] - 1,$$

аппроксимирующую задачу (25′) со вторым порядком и аналогичную схеме (15) для скалярного случая (14). Условие существования нетривиального решения вида (13) у векторного уравнения (25) состоит, как и в примере 8, в том, чтобы обращался в нуль определитель системы, возникающей для определения $u^0 = \begin{bmatrix} v^0 \\ w^0 \end{bmatrix}$.

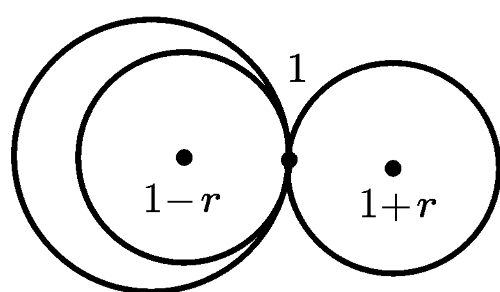


Рис. 25

Приравняв этот определитель нулю, получаем квадратное уравнение относительно $\lambda = \lambda(\alpha)$, из которого находим

$$\begin{aligned} \lambda_1(\alpha) &= 1 + ir\sin\alpha - 2r^2\sin^2\frac{\alpha}{2}, \\ \lambda_2(\alpha) &= 1 - ir\sin\alpha - 2r^2\sin^2\frac{\alpha}{2}. \end{aligned} \tag{30}$$

Эти формулы аналогичны (16), и, как в (17), получим

$$1 - |\lambda_{1,2}(\alpha)|^2 = 4r^2\sin^4\frac{\alpha}{2}(1 - r^2).$$

Спектр, задаваемый формулами (30), лежит в единичном круге при $r \leqslant 1$.

Задачи

1. При каких значениях параметра $\sigma > 0$ разностная схема, аппроксимирующая задачу Коши для уравнения теплопроводности

$$\frac{u_m^{p+1} - u_m^p}{\tau} = \sigma\,\frac{u_{m+1}^{p+1} - 2u_m^{p+1} + u_{m-1}^{p+1}}{h^2} + (1 - \sigma)\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2},$$
$$u_m^0 = \psi(mh), \quad m = 0, \pm 1, ..., \quad p \geqslant 0,$$

удовлетворяет спектральному признаку устойчивости Неймана при любом $r = \tau/h^2 = \text{const}$?

2. Удовлетворяет ли спектральному признаку устойчивости следующая разностная схема:

$$\frac{u_m^{p+1} - u_m^{p-1}}{2\tau} - \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = \varphi(mh, p\tau),$$
$$u_m^0 = \psi(mh), \quad m = 0, \pm 1, ...,$$
$$u_m^1 = \widetilde{\psi}(mh), \quad p \geqslant 0,$$

где

$$\widetilde{\psi}(mh) = u(x,0) + \tau \left.\frac{\partial u(x,0)}{\partial t}\right|_{x=mh} = u(x,0) + \tau \frac{\partial^2 u(x,0)}{\partial x^2} = $$
$$= \psi(mh) + \tau\psi''(mh)?$$

Эта разностная схема аппроксимирует задачу Коши (19) для уравнения теплопроводности с порядком $O(\tau^2 + h^2)$.

§ 4. Принцип замороженных коэффициентов

Здесь мы изложим прием, весьма расширяющий класс нестационарных разностных задач, для исследования которых можно пользоваться спектральным признаком устойчивости. Этот необходимый признак устойчивости, изложенный в § 3 для исследования разностной задачи Коши с постоянными коэффициентами, можно применять и в случае разностной задачи Коши с непрерывными, но не постоянными коэффициентами, а также для задач в ограниченных областях, когда граничные условия задаются не только при $t = 0$, но и на боковых границах. Этим приемом можно пользоваться и для исследования нелинейных задач.

1. Замораживание коэффициентов во внутренних точках. Сформулируем принцип замороженных коэффициентов, воспользовавшись в качестве примера следующей разностной краевой задачей:

$$\frac{u_m^{p+1} - u_m^p}{\tau} - a(x_m, t_p)\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0,$$
$$u_m^0 = \psi(mh), \qquad (1)$$
$$l_1 u_0^{p+1} = 0, \quad l_2 u_M^{p+1} = 0, \quad m = 0, 1, ..., M, \quad Mh = 1, \quad p \geqslant 0.$$

В этой разностной краевой задаче $l_1 u_0^{p+1} = 0$, $l_2 u_M^{p+1} = 0$ — некоторые условия, задаваемые соответственно на левой и правой границах сеточного отрезка.

Выберем произвольную внутреннюю точку $(\widetilde{x}, \widetilde{t})$ области $0 < x < 1$, где рассматривается задача (1), и "заморозим" коэффициенты

задачи (1) в этой точке. Возникающее разностное уравнение с постоянными коэффициентами

$$\frac{u_m^{p+1} - u_m^p}{\tau} - a(\widetilde{x}, \widetilde{t}) \, \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \quad p \geqslant 0, \tag{2}$$

будем рассматривать теперь не при $m = 1, 2, ..., M - 1$, а при всех целочисленных m.

Сформулируем теперь следующий принцип замороженных коэффициентов. *Для устойчивости задачи* (1) *необходимо, чтобы задача Коши для разностного уравнения с постоянными коэффициентами* (2) *удовлетворяла спектральному признаку устойчивости Неймана.*

В обоснование принципа замороженных коэффициентов приведем следующее рассуждение на эвристическом уровне строгости. При измельчении сетки коэффициент $a(x, t)$ в окрестности точки $(\widetilde{x}, \widetilde{t})$ за любое фиксированное число шагов сетки длины h по пространству и длины τ по времени в силу непрерывности функции $a(x, t)$ меняется все меньше и все меньше отличается от значения $a(\widetilde{x}, \widetilde{t})$. Добавим к этому, что расстояние от точки $(\widetilde{x}, \widetilde{t})$ до границ $t = 0$, $x = 0$ и $x = 1$ отрезка, измеренное числом шагов сетки, стремится к бесконечности. Поэтому при мелкой сетке возмущения, наложенные на решение задачи (1) в момент времени $t = \widetilde{t}$ в окрестности точки $x = \widetilde{x}$, развиваются (за малое время) примерно так же, как для задачи (2).

Понятно, что это рассуждение носит общий характер. Оно не зависит от числа пространственных переменных, числа искомых функций и вида разностного уравнения или системы уравнений.

В § 3 мы рассмотрели задачу Коши для уравнения вида (2) и нашли, что для выполнения условия Неймана отношение τ/h шагов сетки должно удовлетворять условию

$$r \leqslant \frac{1}{2a(\widetilde{x}, \widetilde{t})}. \tag{3}$$

Поскольку в силу принципа замороженных коэффициентов для устойчивости задачи (1) это условие должно выполняться при любых $\widetilde{x}$, $\widetilde{t}$, отношение $\tau/h^2 = r$ шагов сетки должно быть подчинено условию

$$r \leqslant \frac{1}{2 \max\limits_{x,t} a(x, t)}. \tag{4}$$

Принцип замороженных коэффициентов позволяет ориентироваться на эвристическом уровне строгости и при исследовании устойчивости нелинейных задач. Поясним это на следующей нелинейной задаче:

$$u_t = (1 + u^2) u_{xx} = 0, \quad 0 < x < 1, \quad 0 < t < T,$$

$$u(x, 0) = \psi(x), \quad 0 < x < 1,$$

$$u(0, t) = \psi^{(1)}(t), \quad u(1, t) = \psi^{(2)}(t).$$

Используем разностную схему

$$L_h u^{(h)} = \begin{cases} \dfrac{u_m^{p+1} - u_m^p}{\tau_p} - [1 + |u_m^p|^2] \dfrac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ u_m^0 = \psi(mh), \quad p = 0, 1, ..., [T/\tau] - 1, \\ u_0^p = \psi^{(1)}(t_p), \quad 0 \leqslant m < M, \\ u_M^p = \psi^{(2)}(t_p), \quad p = 0, 1, ..., [T/\tau]. \end{cases}$$

В ней допускается изменение шага τ_p от слоя к слою. Эта схема позволяет последовательно, слой за слоем, вычислить u_m^1 $(m = 0, 1, ..., M)$, затем u_m^2 $(m = 0, 1, ...)$ и т. д.

Допустим, что мы уже добрались до слоя $t = t_p$, вычислили u_m^p $(m = 0, 1, ..., M)$ и хотим продолжать счет. Как выбрать следующий шаг $\tau = \tau_p$? Можно считать, что нам предстоит найти решение линейного разностного уравнения

$$\frac{u_m^{p+1} - u_m^p}{\tau_p} - a(x_m, t_p) \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0$$

с заданным переменным коэффициентом $a(x_m, t_p) \equiv 1 + |u_m^p|^2$. Действительно, естественно считать, что значения u_m^p близки к значениям $u(mh, t_p)$ гладкого решения $u(x, t)$ дифференциальной задачи. Коэффициент тогда близок к непрерывной функции $a(x, t) = 1 + u^2(x, t)$, которая на протяжении нескольких временны́х шагов мало изменяется.

Применение признака Неймана к уравнению с переменным коэффициентом $a(x_m, t_p)$ дает ограничение (3) на соотношение шагов сетки, необходимое для устойчивости:

$$\frac{\tau_p}{h^2} = r_p \leqslant \frac{1}{2 \max\limits_{x=mh} a(x, t_p)} = \frac{1}{2 \max\limits_m (1 + |u_m^p|^2)}.$$

Отсюда следует рекомендация выбрать очередной шаг τ_p из условия

$$\tau_p \leqslant \frac{h^2}{2 \max\limits_m (1 + |u_m^p|^2)}.$$

Численный эксперимент на компьютере подтверждает правильность этих эвристических рассуждений.

Если необходимое условие устойчивости, полученное путем рассмотрения задачи Коши с замороженными в произвольной точке области коэффициентами, окажется нарушенным, то устойчивости нельзя ожидать ни при каком задании граничных условий. Подчеркнем, однако, что принцип замороженных коэффициентов не учитывает влияния граничных условий. В случае выполнения необходимого условия устойчивости, вытекающего из принципа замороженных коэффициентов, устойчивость может иметь место при одних

граничных условиях и не иметь места при других. Теперь изложим признак Бабенко–Гельфанда, учитывающий влияние границ в случае задачи на отрезке.

2. Признак Бабенко–Гельфанда. При рассмотрении задачи (1) мы полагали, что возмущения, сообщенные решению задачи (1) в окрестности произвольной внутренней точки $(\widetilde{x}, \widetilde{t})$, при мелкой сетке развиваются примерно так же, как такие же возмущения, сообщенные решению задачи Коши (2) с замороженными в точке $(\widetilde{x}, \widetilde{t})$ коэффициентами. В обоснование этого принципа мы принимали во внимание, что расстояния от внутренней точки $(\widetilde{x}, \widetilde{t})$ до границ, измеренные числом шагов сетки, при измельчении сетки неограниченно возрастают. Но если точка $(\widetilde{x}, \widetilde{t})$ лежит на боковой границе $x = 0$ или $x = 1$, то это эвристическое рассуждение теряет убедительность. Пусть, например, $\widetilde{x} = 0$. Тогда расстояние от точки $(\widetilde{x}, \widetilde{t})$ до любой фиксированной точки $x > 0$ (в частности, до правого конца отрезка $x = 1$), измеренное числом шагов сетки, при $h \to 0$ по-прежнему неограниченно возрастает, но число шагов до левого конца $x = 0$ не меняется и остается равным нулю.

Поэтому возмущение решения задачи (1) вблизи левой границы $x = 0$ за малое время должно развиваться подобно возмущению решения задачи

$$\frac{u_m^{p+1} - u_m^p}{\tau} - a(0, \widetilde{t})\, \frac{u_{m+1}^{p+1} - 2u_m^p + u_{m-1}^p}{h^2} = 0, \quad m = 1, 2, \ldots,$$
$$l_1 u_0^{p+1} = 0.$$

Эта задача получилась из исходной задачи (1) при замораживании коэффициента $a(x, t)$ в левом конце отрезка $0 \leqslant x \leqslant 1$ и одновременном удалении правой границы $b + \infty$. Задачу (4) естественно рассматривать только на тех функциях $u^p = \{u_0^p, u_1^p, \ldots\}$, для которых

$$u_m^p \to 0, \quad m \to +\infty.$$

Только в этом случае возмущение сосредоточено вблизи границы $x = 0$, и только относительно возмущений такого вида задача (1) и задача (4) вблизи левой границы $x = 0$ сходны между собой.

Точно так же развитие возмущений решения задачи (1) вблизи правой границы $x = 1$ должно быть похоже на развитие таких же возмущений для задачи

$$\begin{gathered} \frac{u_m^{p+1} - u_m^p}{\tau} - a(1, \widetilde{t})\, \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\ m = \ldots, -1, 0, 1, \ldots, M - 1, \quad p \geqslant 0, \\ l_2 u_M^{p+1} = 0 \end{gathered} \tag{5}$$

с одной только правой границей. Эта задача возникла из исходной задачи (1) при замораживании коэффициента $a(x, t)$ в правом конце

$x=1$ и при удалении левой границы $b-\infty$. Задачу (5) надо рассматривать на сеточных функциях $u^p=\{\ldots,u_{-1}^p,u_0^p,u_1^p,\ldots,u_M^p\}$, удовлетворяющих условию $u_m^p\to 0$ при $m\to-\infty$.

Задачи (2), (4), (5) проще исходной задачи (1) в том отношении, что при фиксированном $r=\tau/h^2$ они не зависят от h и являются задачами с постоянными коэффициентами.

Таким образом, процедура исследования устойчивости, учитывающая влияние границ, применительно к задаче (1) состоит в следующем. Надо составить вспомогательные задачи (2), (4), (5). Для каждой из этих трех задач, не зависящих от h, надо найти все те числа λ (собственные числа оператора перехода от u^p к u^{p+1}), при которых существуют решения вида

$$u_m^p=\lambda^p u_m^0.$$

При этом в случае задачи (2) функция $u^0=\{u_m^0\}$ $(m=0,\pm 1,\ldots)$ должна быть ограничена. В случае задачи (4) для $u^0=\{u_m^0\}$ $(m\geqslant 0)$ должно быть $u_m^0\to 0$ $(m\to+\infty)$, а в случае задачи (5) должно быть $u^0=\{u_m\}$ $(m\leqslant M)$, $u_m^0\to 0$ $(m\to-\infty)$.

Для устойчивости задачи (1) совокупность собственных чисел каждой из трех задач (2), (4), (5) должна лежать в единичном круге $|\lambda|<1$. (Задача (2) рассматривается при каждом фиксированном $\widetilde{x}$ $(0<\widetilde{x}<1)$.)

Продолжим рассмотрение задачи (1). Будем считать в дальнейшем, что $a(x,t)\equiv 1$, и вычислим спектры для всех трех задач (2), (4), (5) при различных краевых условиях $l_1u_0^{p+1}=0$, $l_2u_M^{p+1}=0$.

Подставляя решение $u_m^p=\lambda^p u_m$ в разностное уравнение (2), получаем

$$(\lambda-1)u_m-r(u_{m+1}-2u_m+u_{m-1})=0,\quad r=\tau/h^2,$$

или

$$u_{m+1}-\frac{-2r+1-\lambda}{r}u_m+u_{m-1}=0. \tag{6}$$

Это — обыкновенное разностное уравнение второго порядка.

Чтобы написать общее решение уравнения (6), составим характеристическое уравнение

$$q^2-\frac{-2r+1-\lambda}{r}q+1=0. \tag{7}$$

Если q — корень этого уравнения, то сеточная функция

$$u_m=\lambda^p q^m$$

есть одно из решений уравнения

$$\frac{u^{p+1}-u_m^p}{\tau}-\frac{u_{m+1}^p-2u_m^p+u_{m-1}^p}{h^2}=0.$$

Если $|q| = 1$, т. е. $q = e^{i\alpha}$, то ограниченная при $m \to +\infty$ и при $m \to -\infty$ сеточная функция

$$u_m^p = \lambda^p e^{i\alpha m},$$

как мы видели в § 3, является решением при

$$\lambda = 1 - 4r\sin^2(\alpha/2), \quad 0 \leqslant \alpha < 2\pi.$$

Эти $\lambda = \lambda(\alpha)$ заполняют отрезок $1 - 4r \leqslant \lambda \leqslant 1$ на вещественной оси. Этот отрезок и есть спектр задачи (2). Собственных значений λ, не лежащих на этом отрезке, задача (2) не имеет, так как в случае отсутствия у характеристического уравнения (7) корня q, по модулю равного единице, задача (6) не имеет ограниченного при $m \to \pm\infty$ решения.

Если λ не лежит на отрезке $1 - 4r \leqslant \lambda \leqslant 1$, то оба корня характеристического уравнения (7) отличны по модулю от единицы, но их произведение равно свободному члену квадратного уравнения (7), т. е. единице. Поэтому среди корней уравнения (7) один по модулю больше, а другой меньше единицы. Пусть для определенности $|q_1(\lambda)| < 1$, $|q_2| > 1$. Тогда общее решение уравнения (6), убывающее по модулю при $m \to +\infty$, имеет вид

$$u_m = cq_1^m, \quad q_1 = q_1(\lambda) < 1,$$

а общее решение уравнения (6), стремящееся к нулю при $m \to -\infty$, имеет вид

$$u_m = cq_2^m, \quad q_2 = q_2(\lambda) > 1.$$

Для определения собственных значений задачи (4) надо подставить $u_m = cq_1^m$ в левое граничное условие $l_1 u = 0$ и найти те λ, при которых оно выполняется. Это и будут все собственные значения задачи (4). Если, например, $l_1 u_0 \equiv u_0 = 0$, то условие $cq_1^0 = 0$ не выполняется ни при каком $c \neq 0$, так что собственных значений нет. Если $l_1 u_0 = u_1 - u_0 = 0$, то условие $c(q_1 - q_1^0) = c(q_1 - 1) = 0$ в силу $q_1 \neq 1$ приводит к $c = 0$, так что собственных значений опять нет. В случае $l_1 u_0 = 2u_1 - u_0 = 0$ условие $c(2q_1 - q_1^0) = 0$ выполняется при $c \neq 0$, если $q_1 = 1/2 < 1$.

Из уравнения (7) находим, что при $q_1 = 1/2$ число λ есть

$$\lambda = 1 + r\Big(q_1 - 2 + \frac{1}{q_1}\Big) = 1 + r\,\frac{1 - 4 + 4}{2} = 1 + \frac{r}{2}.$$

Это и есть единственное собственное значение задачи (4). Оно лежит вне единичного круга устойчивости. Аналогично вычисляются собственные значения задачи (5). Они получаются из уравнения

$$l_2 u_M = 0$$

при

$$u_m = q_2^m, \quad q_2 = q_2(\lambda), \quad m = M, M-1, \ldots$$

Рассмотрим в качестве еще одного примера разностную схему

$$\begin{gathered} \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad p = 0, 1, \ldots, \left[\frac{T}{\tau}\right] - 1, \\ m = 0, 1, \ldots, M-1, \quad Mh = 1, \\ u_m^0 = \psi(mh), \quad u_M^{p+1} = 0, \end{gathered} \tag{8}$$

аппроксимирующую задачу

$$\begin{gathered} \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad 0 \leqslant x \leqslant 1, \quad 0 < t < T, \\ u(x, 0) = \psi(x), \quad u(1, t) = 0. \end{gathered}$$

Применим для исследования ее устойчивости признак Бабенко–Гельфанда. Сопоставим по схеме (8) три задачи: задачу без боковых границ

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad m = 0, \pm 1, \ldots, \tag{9}$$

задачу с одной только левой боковой границей

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad m = 0, 1, \ldots, \tag{10}$$

и задачу с одной только правой боковой границей

$$\begin{gathered} \frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_m^p}{h} = 0, \quad m = M-1, M-2, \ldots, \\ u_M^{p+1} = 0. \end{gathered} \tag{11}$$

В случае задачи (10) — с одной только левой боковой границей — граничного условия нет, так как его не было в исходной задаче (8).

Надо найти совокупность собственных чисел всех трех операторов перехода от u^p к u^{p+1}, соответствующих каждой из трех вспомогательных задач (9)–(11), и выяснить, при каких условиях все они лежат в круге $|\lambda| \leqslant 1$.

Решение вида

$$u_m^p = \lambda^p u_m^0$$

при подстановке в разностное уравнение

$$u_m^{p+1} = (1-r)u_m^p + r u_{m+1}^p, \quad r = \tau/h,$$

приводит к следующему обыкновенному разностному уравнению первого порядка для собственной функции:

$$(\lambda - 1 + r)u_m - r u_{m+1} = 0. \tag{12}$$

Соответствующее характеристическое уравнение

$$\lambda - 1 + r - rq = 0 \tag{13}$$

дает связь между λ и q. Общее решение уравнения (12) есть

$$u_m = cq^m = c\left(\frac{\lambda - 1 + r}{r}\right)^m, \quad m = 0, \pm 1, \ldots$$

При $|q| = 1$, $q = e^{i\alpha}$ $(0 \leqslant \alpha \leqslant 2\pi)$

$$\lambda = 1 - r + re^{i\alpha}.$$

Точка $\lambda = \lambda(\alpha)$ пробегает окружность с центром в точке $1 - r$ и ради-

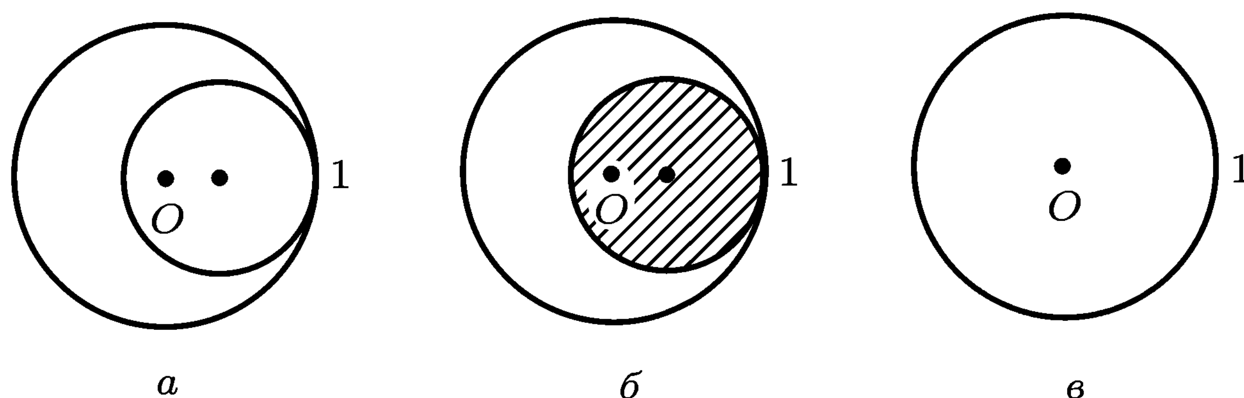


Рис. 26

усом r. Это и есть собственные значения задачи (9) (рис. 26, *а*). Убывающее при $m \to +\infty$ нетривиальное решение задачи (10) существует при любом q $(|q| < 1)$.

Соответствующие $\lambda = 1 - r + rq$ заполняют, очевидно, всю внутренность круга, ограниченного окружностью $\lambda = (1 - r) + re^{i\alpha}$ (рис. 26, *б*).

Наконец, решения задачи (11) $u_m^p = \lambda^p u_m$, убывающие при $m \to \to -\infty$, должны иметь вид

$$u_m^p = c\lambda^p q^m, \quad |q| > 1,$$

где λ, q связаны равенством (13).

Из граничного условия $u_M^p = 0$ следует, что нетривиальное реше-

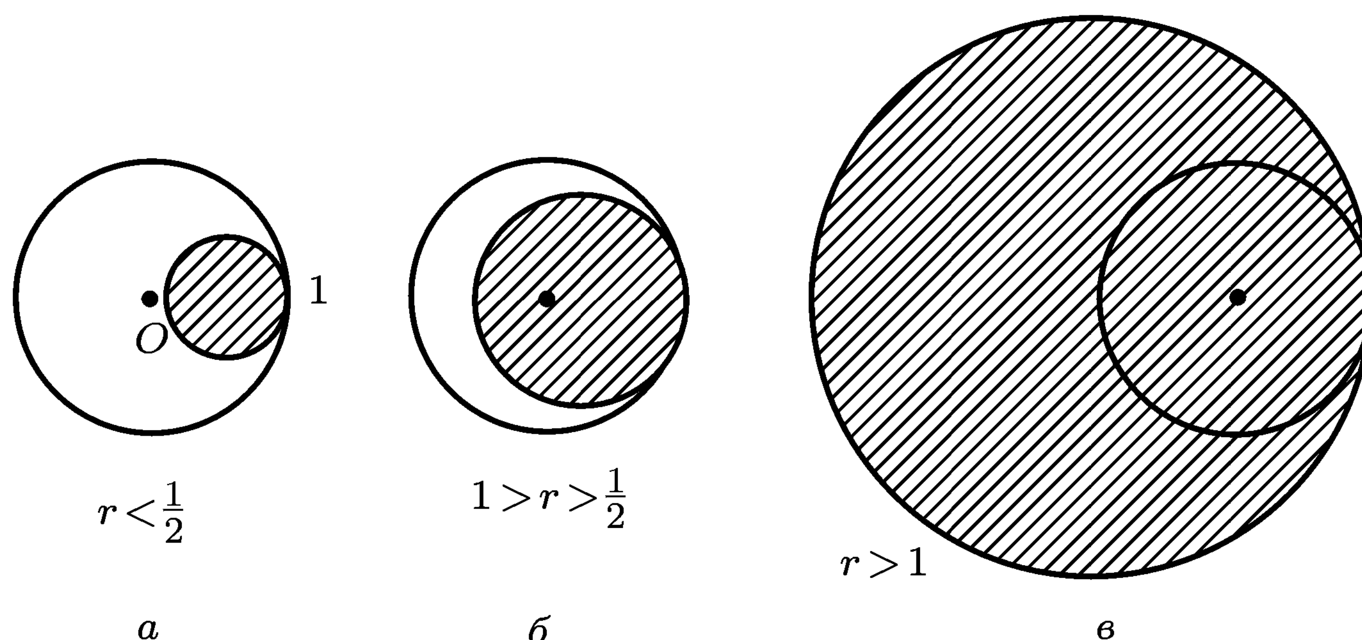


Рис. 27

ние существует только при $\lambda = \lambda(q) = 0$, т. е. при $q = (r - 1)/r$. Эта

величина q по модулю больше единицы в случае выполнения одного из неравенств, а именно, $(r-1)/r > 1$ или $(r-1)/r < -1$. Первое неравенство решений не имеет. Решение второго имеет вид $r < 1/2$.

Итак, при $r < 1/2$ задача (10) имеет собственное значение $\lambda = 0$ (рис. 26, *в*). На рис. 27 изображены объединения собственных значений всех трех задач соответственно для случаев $r < 1/2$, $1/2 < r < 1$, $r > 1$.

Ясно, что объединение собственных значений всех трех задач лежит в круге $|\lambda| < 1 + c\tau$, где c не зависит от τ в том и только том случае, если $r \leqslant 1$.

Изложенный здесь признак устойчивости нестационарных разностных задач на отрезке, учитывающий влияние граничных условий, применим и в случае краевых задач на отрезке для систем разностных уравнений. В этом случае естественные на первый взгляд схемы, удовлетворяющие признаку Неймана, часто оказываются неустойчивыми из-за неудачной аппроксимации граничных условий, и важно уметь подбирать схемы, свободные от этого недостатка.

В [5, 6] спектральный признак Бабенко–Гельфанда обсуждается с более общей точки зрения — на основе введенного там понятия *спектра семейства операторов.* В частности, С.К. Годуновым и В.С. Рябеньким строго показано, что его выполнение необходимо для устойчивости и что в случае его выполнения устойчивость не может грубо нарушаться.

Задачи

1. Выяснить условия выполнения спектрального признака устойчивости для разностной схемы

$$\frac{u_m^{p+1} - u_m^p}{\tau} - \frac{u_{m+1}^p - u_{m-1}^p}{2h} - \frac{\tau}{2}\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0,$$

$$p \geqslant 0, \quad m = 1, 2, \ldots, M-1, \quad Mh = 1,$$

$$u_m^0 = \psi(mh), \quad m = 0, 1, \ldots, M,$$

$$\frac{u_0^{p+1} - u_0^p}{\tau} - \frac{u_1^p - u_0^p}{h} = 0,$$

$$u_M^{p+1} = 0, \quad p = 0, 1, \ldots, [T/\tau] - 1,$$

аппроксимирующей дифференциальную задачу

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad 0 \leqslant x \leqslant 1, \quad 0 \leqslant t \leqslant T,$$

$$u(x, 0) = \varphi(x), \quad u(1, t) = 0$$

на гладком решении $u(x,t)$ со вторым порядком относительно h.

Ответ. $\tau/h \leqslant 1$.

2. Для построения разностной схемы, аппроксимирующей следующую краевую задачу для гиперболической системы дифференциальных уравнений:

$$\frac{\partial v}{\partial t} = \frac{\partial w}{\partial x}, \quad \frac{\partial w}{\partial t} = \frac{\partial v}{\partial x}, \quad 0 \leqslant x \leqslant 1, \quad 0 \leqslant t \leqslant T,$$

$$v(x,0) = \psi^{(1)}(x),$$
$$w(x,0) = \psi^{(2)}(x),$$
$$v(0,t) = w(1,t) = 0,$$

положим $u(x,t) = \begin{bmatrix} v(x,t) \\ w(x,t) \end{bmatrix}$, $\psi(x) = \begin{bmatrix} \psi^{(1)}(x) \\ \psi^{(2)}(x) \end{bmatrix}$ и запишем ее в матричной форме

$$\frac{\partial u}{\partial t} - A\frac{\partial u}{\partial x} = 0,$$
$$u(x,0) = \psi(x),$$
$$v(0,t) = w(1,t) = 0,$$

где $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Выберем сетку $(x_m, t_p) = (mh, p\tau)$, $h = M^{-1}$, M — натуральное число. Положим

$$\frac{u_m^{p+1} - u_m^p}{\tau} - A\,\frac{u_{m+1}^{p+1} - u_{m-1}^p}{2h} - \frac{A^2\tau}{2}\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0,$$
$$m = 1, 2, \ldots, M-1, \quad p \geqslant 0,$$
$$u_m^0 = \psi(mh), \quad v_0^{p+1} = w_M^{p+1} = 0.$$

Для завершения построения схемы надо задать дополнительные граничные условия на левой и правой боковых границах. Заметив, что при любых α, β из системы дифференциальных уравнений следуют равенства

$$\left.\frac{\partial(v+\alpha w)}{\partial t} - \frac{\partial(w+\alpha v)}{\partial x}\right|_{x=0} = 0,$$
$$\left.\frac{\partial(v+\beta w)}{\partial t} - \frac{\partial(w+\beta v)}{\partial x}\right|_{x=1} = 0,$$

зададим дополнительные разностные краевые условия, положив

$$\frac{(v_0^{p+1} + \alpha w_0^{p+1}) - (v_0^p + \alpha w_0^p)}{\tau} - \frac{(w_1^p + \alpha v_1^p) - (w_0^p + \alpha v_0^p)}{h} = 0,$$
$$\frac{(v_M^{p+1} + \beta w_M^{p+1}) - (v_M^p + \beta w_M^p)}{\tau} - \frac{(w_M^p + \beta v_M^p) - (w_{M-1}^p - \beta v_{M-1}^p)}{h} = 0.$$

При условии $r = \tau/h \leqslant 1$ показать, что:

а) если $\alpha = 1$, $\beta = -1$, то спектральный признак устойчивости выполнен;

б) если $\alpha = -1$, то независимо от выбора числа β спектральный признак устойчивости не выполнен;

в) найти условия, которым должны удовлетворять α, β, чтобы выполнялся спектральный признак устойчивости, учитывающий влияние граничных условий.

§ 5. Явные и неявные разностные схемы для уравнения теплопроводности

Рассмотрим следующую задачу для модельного уравнения теплопроводности:

$$
\begin{gathered}
\frac{\partial u}{\partial t} - a(x,t)\,\frac{\partial^2 u}{\partial x^2} = 0, \quad a(x,t) > 0, \quad 0 \leqslant x \leqslant 1, \\
u(x,0) = \psi(x), \\
u|_{x=0} = \varphi_{\text{лев.}}(t), \quad u|_{x=1} = \varphi_{\text{прав.}}(t).
\end{gathered} \tag{1}
$$

Для численного решения этой задачи рассмотрим и сравним две разностные схемы — явную и неявную. При этом выяснится, что в некоторых случаях неявная схема обладает преимуществами перед явной, несмотря на то, что сам алгоритм вычислений по явной схеме проще. Это преимущество связано с ее устойчивостью при любом соотношении шагов сетки по времени и по пространству.

1. Явная разностная схема. Воспользуемся следующим разностным аналогом (см. рис. 18):

$$
\begin{gathered}
\frac{u_m^{p+1} - u_m^p}{\tau_p} - a(x_m, t_p)\,\frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2} = 0, \\
u_m^0 = \psi_m, \quad m = 0, 1, ..., M, \quad Mh = 1, \\
u_0^{p+1} = \psi_{\text{лев.}}(t_{p+1}), \quad u_M^{p+1} = \varphi_{\text{прав.}}(t_{p+1}), \\
t_0 = 0, \quad t_p = \tau_0 + \tau_1 + ... + \tau_{p-1}, \quad p = 1, 2, ...
\end{gathered} \tag{2}
$$

Пусть решение u_m^k $(m = 0, 1, ..., M)$ уже найдено при $k = 0, 1, ..., p$. Тогда вычисление u_m^{p+1} на временно́м слое $t = t_{p+1} = t_p + \tau_p$ осуществляется в силу (2) по явной формуле

$$
u_m^{p+1} = u_m^p + \frac{\tau_p}{h^2}\,a(x_m, t_p)(u_{m+1}^p - 2u_m^p + u_{m-1}^p), \tag{3}
$$

$$
m = 1, 2, ..., M - 1.
$$

Поэтому разностная схема (2) называется *явной разностной схемой.*

В силу принципа замороженных коэффициентов устойчивости можно ожидать лишь в том случае, если

$$
\tau_p \leqslant \frac{h^2}{2 \max\limits_m a(x_m, t_p)}. \tag{4}
$$

Отсюда следует, что в случае больших значений $a(x_m, t_p)$ в окрестности какой-нибудь точки $(\widetilde{x}, \widetilde{t})$ для вычисления решения на слое $t = t_{p+1}$ придется воспользоваться очень маленьким значением τ, так что продвижение до заданного $t = T$ может потребовать очень большого числа шагов по времени и оказаться практически слишком трудоемким.

Заметим, что это не связано с физической сутью дела. В самом деле, большое значение коэффициента теплопроводности $a(x_m, t_p)$ в окрестности точки $(\widetilde{x}, \widetilde{t})$ физически означает просто, что эту окрестность можно "вырезать" из теплопроводного стержня, не изменив картину распространения тепла: можно считать, что этот кусочек стержня сделан из материала с нулевой теплоемкостью.

2. Неявная разностная схема. Вместо схемы (2) воспользуемся неявной разностной схемой

$$\begin{gathered}\frac{u_m^{p+1}-u_m^p}{\tau}-a(x_m,t_p)\,\frac{u_{m+1}^{p+1}-2u_m^{p+1}+u_{m-1}^{p+1}}{h^2}=0,\\ m=1,2,\ldots,M-1,\\ u_m^0=\psi_m,\quad m=0,1,\ldots,M,\\ u_0^{p+1}=\psi_{\text{лев.}}(t_{p+1}),\quad u_M^{p+1}=\psi_{\text{прав.}}(t_{p+1}),\quad p\geqslant 0.\end{gathered}\tag{5}$$

Пусть решение на слое $t = t_p$ уже найдено, так что значения u_m^p $(m = 0, 1, ..., M)$ уже вычислены. Для вычисления решения u_m^{p+1} $(m = 0, 1, ..., M)$ надо решить систему (5) вида

$$\begin{gathered}u_0=\varphi_{\text{лев.}}(t_{p+1}),\\ \alpha_m u_{m-1}+\beta_m u_m+\gamma_m u_{m+1}=f_m,\quad m=1,2,\ldots,M-1,\\ u_M=\psi_{\text{прав.}}(t_{p+1}),\end{gathered}$$

где

$$\begin{aligned}&\alpha_m=\gamma_m=+\frac{a(x_m,t_p)}{h^2}\tau_p,\\ &\beta_m=\frac{2a(x_m,t_p)}{h^2}\tau_p+1,\quad f_m=u_m^p.\end{aligned}$$

Очевидно, что

$$|\beta_m|>|\alpha_m|+|\gamma_m|+\delta,\quad \delta=1>0,$$

так что выполнены условия применимости метода прогонки.

В § 3 (см. пример 5) мы рассмотрели неявную разностную схему для уравнения теплопроводности с постоянным коэффициентом и установили, что спектральный признак устойчивости выполнен при любом отношении $\tau/h^2 = r$ шагов сетки по времени и пространству. В силу принципа замороженных коэффициентов спектральный признак не накладывает ограничений на шаги сетки и в случае переменного коэффициента теплопроводности. Это делает неявную схему пригодной и в тех случаях, когда $a(x, t)$ в отдельных местах принимает очень большие значения.

В заключение отметим, что условие устойчивости (4) для явной схемы (2) и абсолютная устойчивость схемы (5) установлены нами на эвристическом уровне строгости, поскольку мы опирались на принцип замороженных коэффициентов, носящий эвристический ха-

рактер. Однако наши выводы о свойствах явной и неявной разностных схем (2), (5) действительно справедливы и могут быть строго доказаны.

Задача

Пусть коэффициент теплопроводности в задаче (1) задается формулой $a = 1 + u^2$, так что задача (1) становится нелинейной.

а) Предложить явную и неявную разностные схемы для этой задачи.

б) Рассмотреть явную схему

$$\frac{u_m^{p+1} - u_m^p}{\tau_p} = [1 + (u_m^p)^2]\, \frac{u_{m+1}^p - 2u_m^p + u_{m-1}^p}{h^2},$$

$$m = 1, 2, ..., M - 1,$$

$$u_0^p = \varphi_{\text{лев.}}^p, \quad u_m^0 = \psi_m, \quad u_M^p = \varphi_{\text{прав.}}^p, \quad p = 0, 1, ...$$

Как выбирать τ_p по значениям u_m^p решения на слое p?

в) Рассмотреть неявную разностную схему на основе следующего разностного уравнения:

$$\frac{u_m^{p+1} - u_m^p}{\tau} = [1 + (u_m^{p+1})^2]\, \frac{u_{m+1}^{p+1} - 2u_m^{p+1} - u_{m-1}^{p+1}}{h^2}.$$

Как следует усовершенствовать это уравнение, чтобы при переходе от u_m^p $(m = 0, 1, ..., M)$ к u_m^{p+1} $(m = 0, 1, ..., M)$ можно было воспользоваться методом прогонки?

ГЛАВА 10
ПОНЯТИЕ О РАЗРЫВНЫХ РЕШЕНИЯХ И СПОСОБАХ ИХ ВЫЧИСЛЕНИЯ

Такие дифференциальные уравнения математической физики, как уравнения газовой динамики, упругости, теплопроводности и др., являются следствиями физических законов сохранения величин (массы, импульса, энергии и т. д.) [21], записываемых в некоторой интегральной форме. Эти дифференциальные уравнения равносильны исходным интегральным соотношениям только в том случае, если искомые физические поля (т. е. искомые функции) дифференцируемы.

Во всех рассмотренных до сих пор примерах мы предполагали, что существуют достаточно гладкие решения дифференциальных краевых задач, а в основу построения разностных схем клали приближенную замену производных в дифференциальном уравнении разностными отношениями. Однако дифференцируемых функций недостаточно для описания многих важных процессов физики. Так, например, физические эксперименты показывают, что распределения

давления, плотности и скорости в сверхзвуковом течении невязкого газа описываются функциями, имеющими скачки. Скачки (разрывы) могут возникать с течением времени при гладких начальных данных.

Интегральная форма записи законов сохранения имеет смысл и для разрывных решений, так как разрывные функции можно интегрировать. Однако теряется возможность перейти к равносильной дифференциальной постановке задачи, так как разрывные функции нельзя дифференцировать. Вместе с тем теряются удобства, которыми мы пользовались при исследовании свойств решений дифференциальных краевых задач и при построении разностных схем. Поэтому прежде чем переходить к построению алгоритмов вычисления разрывных решений задач, сформулированных в терминах интегральных законов сохранения, полезно так обобщить понятие решения дифференциальной краевой задачи, чтобы она сохраняла смысл и оставалась равносильной исходному интегральному закону сохранения даже в случае разрывных решений.

Мы изложим способ перехода от постановки задачи в терминах интегрального закона сохранения к равносильной дифференциальной краевой задаче, имеющей лишь обобщенное (разрывное) решение, а также способы вычисления обобщенного решения на примере следующей задачи.

В полосе $0 \leqslant t \leqslant T$ найти функцию $u(x,t)$, для которой при произвольном замкнутом контуре Γ выполнено равенство вида

$$\int \frac{u^k}{k}\,dx - \frac{u^{k+1}}{k+1}\,dt \underset{\Gamma}{\equiv} 0, \qquad \text{(А)}$$

где k — некоторое фиксированное натуральное число, и которая удовлетворяет начальному условию

$$u(x,0) = \psi(x), \quad -\infty < x < \infty. \qquad \text{(Б)}$$

Левую часть уравнения (А) можно понимать как поток вектора

$$\Phi(x,t) = \begin{bmatrix} u^k/k \\ u^{k+1}/(k+1) \end{bmatrix}$$

через контур Γ. Требование, чтобы поток этого вектора через произвольный контур Γ был равен нулю, будем понимать как некоторый закон сохранения, записанный в интегральной форме.

Задача (А), (Б) является простейшей среди обладающих свойством возникновения разрывных решений из гладких начальных данных. Поэтому она может служить модельной для понимания способов решения подобных задач в газовой динамике.

§ 1. Дифференциальная формулировка интегрального закона сохранения

1. Дифференциальное уравнение в случае гладких решений. Предположим сначала, что задача (А), (Б) имеет решение $u(x,t)$, непрерывное вместе со своими производными всюду в полосе $0 \leqslant t \leqslant T$. Покажем, что в таком случае она равносильна следующей задаче Коши:

$$\begin{gathered} \frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = 0, \quad 0 < t < T, \quad -\infty < x < \infty, \\ u(x,0) = \psi(x). \end{gathered} \tag{1}$$

Уравнение (1) называется *уравнением Хопфа.* Для этого напомним формулу Остроградского–Грина. Пусть D — произвольная область с границей Γ на плоскости Oxt, и пусть функции $\Phi_1(x,t)$, $\Phi_2(x,t)$ имеют в области D непрерывные вплоть до границы Γ частные производные. Тогда справедлива следующая *формула Остроградского–Грина*:

$$\iint\limits_D \left(\frac{\partial \Phi_1}{\partial t} + \frac{\partial \Phi_2}{\partial x}\right) dx\,dt = \int\limits_\Gamma \Phi_1 dx - \Phi_2\,dt. \tag{2}$$

Это тождество означает, что интеграл от дивергенции $\partial\Phi_1/\partial t + \partial\Phi_2/\partial x$ векторного поля $\Phi = \begin{bmatrix} \Phi_1 \\ \Phi_2 \end{bmatrix}$ по области D равен потоку вектора Φ через границу Γ этой области. Пользуясь этой формулой, можно написать

$$\int\limits_\Gamma \frac{u^k}{k}\,dx - \frac{u^{k+1}}{k+1}\,dt = \iint\limits_D \left[\frac{\partial}{\partial t}\left(\frac{u^k}{k}\right) + \frac{\partial}{\partial x}\left(\frac{u^{k+1}}{k+1}\right)\right] dx\,dt. \tag{3}$$

Отсюда видно, что если гладкая функция $u(x,t)$ удовлетворяет уравнению (1), то выполнено и равенство (А). В самом деле, если выполнено (1), то также

$$u^{k-1}\left(\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x}\right) \equiv \frac{\partial}{\partial t}\left(\frac{u^k}{k}\right) + \frac{\partial}{\partial x}\left(\frac{u^{k+1}}{k+1}\right) = 0. \tag{4}$$

Поэтому правая часть (3) обращается в нуль. Справедливо и обратное: если гладкая функция $u(x,t)$ удовлетворяет интегральному закону сохранения (А), то в каждой точке (x_0,t_0) полосы $0 < t < T$ выполнено уравнение (4), а вместе с тем и (1). Для доказательства предположим противное, и пусть для определенности в некоторой точке (x_0,t_0)

$$\frac{\partial}{\partial t}\left(\frac{u^k}{k}\right) + \frac{\partial}{\partial x}\left(\frac{u^{k+1}}{k+1}\right)\Bigg|_{\substack{x=x_0\\ t=t_0}} > 0.$$

Тогда в силу непрерывности можно найти столь малый круг D с границей Γ и с центром в точке (x_0, t_0), всюду в котором

$$\frac{\partial}{\partial t}\left(\frac{u^k}{k}\right) + \frac{\partial}{\partial x}\left(\frac{u^{k+1}}{k+1}\right) > 0.$$

Получим

$$0 = \int\limits_{\Gamma} \frac{u^k}{k}\,dx - \frac{u^{k+1}}{k+1}\,dt = \iint\limits_{D} \left[\frac{\partial}{\partial t}\left(\frac{u^k}{k}\right) + \frac{\partial}{\partial x}\left(\frac{u^{k+1}}{k+1}\right)\right] dx\,dt > 0.$$

Возникшее противоречие $0 > 0$ доказывает сделанное утверждение: для гладких функций $u(x,t)$ задача (А), (Б) и задача Коши (1) равносильны.

2. Механизм возникновения разрывов. Предположим сначала, что задача (1) имеет гладкое решение $u(x,t)$. Введем линии $x = x(t)$, определяемые уравнением

$$\frac{dx}{dt} = u(x,t). \tag{5}$$

Эти линии называются *характеристиками* уравнения $u_t + uu_x = 0$.

Вдоль каждой характеристики $x = x(t)$ решение $u(x,t)$ можно считать функцией, зависящей только от t:

$$u(x,t) = u[x(t), t] = u(t).$$

Тогда

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = 0.$$

Поэтому вдоль характеристики решение постоянно: $u(x,t) = \mathrm{const}$. Но в силу уравнения (5) из $u = \mathrm{const}$ следует, что характеристика есть прямая линия

$$x = ut + x_0.$$

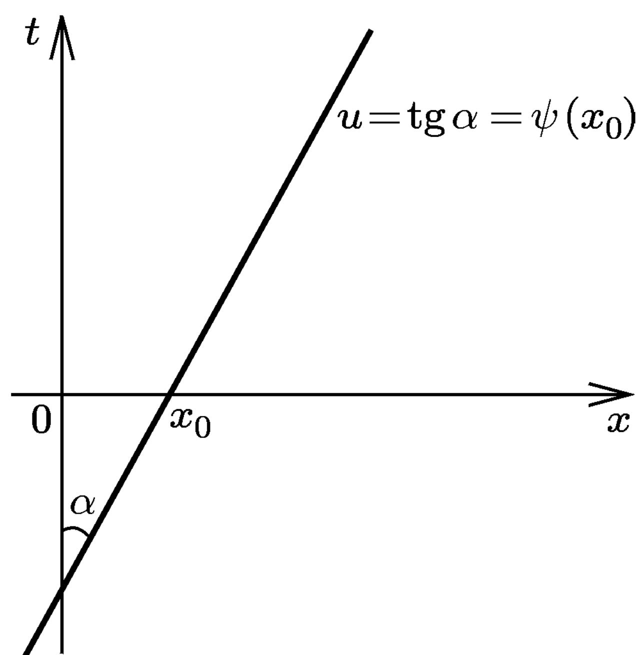


Рис. 28

Здесь x_0 — абсцисса точки $(x_0, 0)$, из которой выходит характеристика, а $u = \psi(x_0)$ — угловой коэффициент ее наклона к оси Ot. Заданием начальной функции $u(x,0) = \psi(x)$, таким образом, наглядно определяется и картина характеристик, и значения решения $u(x,t)$ в каждой точке полуплоскости $t > 0$ (рис. 28).

Заметим сразу же, что в предположении существования гладкого решения $u(x,t)$ характеристики не могут пересекаться, так как каждая характеристика приносила бы в

точку пересечения свое значение решения и решение не было бы однозначной функцией. При монотонно возрастающей функции $\psi(x)$ с ростом x_0 угол α увеличивается, характеристики не пересекаются (рис. 29). Но в случае убывания функции $\psi(x)$ характеристики сходятся и пересечения неизбежны независимо от гладкости функции

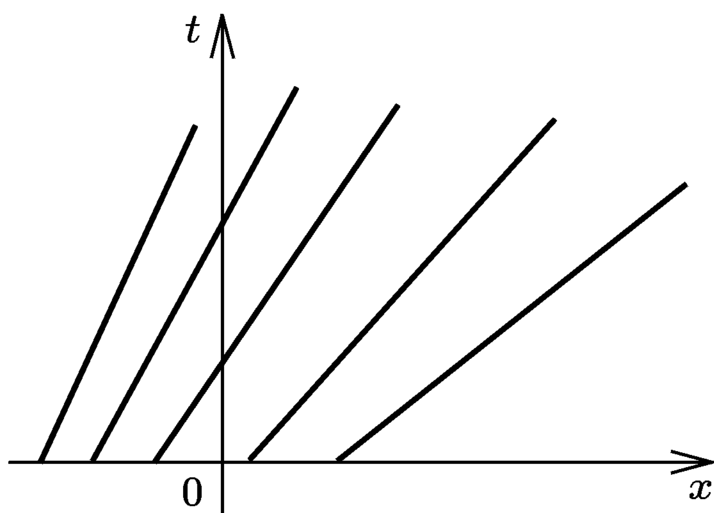


Рис. 29

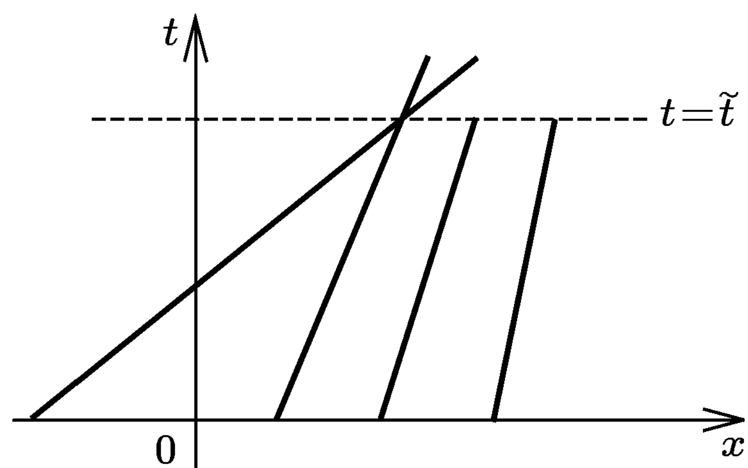


Рис. 30

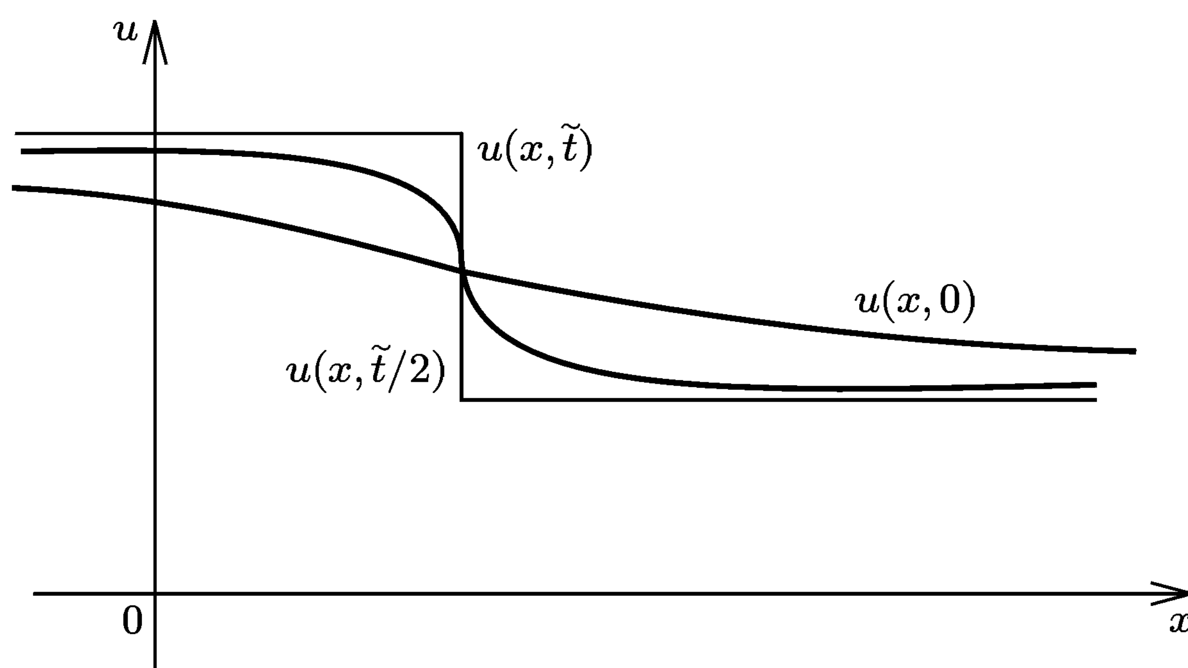


Рис. 31

$u(x,t)$. Гладкое решение задачи (1) перестает существовать с момента $t = \widetilde{t}$, когда хотя бы две характеристики пересекутся (рис. 30); графики функции $u(x,t) = u$ при $t = 0$, $t = \widetilde{t}/2$, $t = \widetilde{t}$ изображены на рис. 31.

3. Условие на линии разрыва решения. Пусть внутри области, где отыскивается решение, имеется линия $x = x(t)$ на которой решение $u(x,t)$ задачи (А), (Б) терпит разрыв первого рода. Пусть при приближении к этой линии слева или справа получаем на ней соответственно

$$u(x,t) = u_{\text{лев.}}(x,t), \quad u(x,t) = u_{\text{прав.}}(x,t).$$

Оказывается, что значения $u_{\text{лев.}}(x,t)$, $u_{\text{прав.}}(x,t)$ и скорость движе-

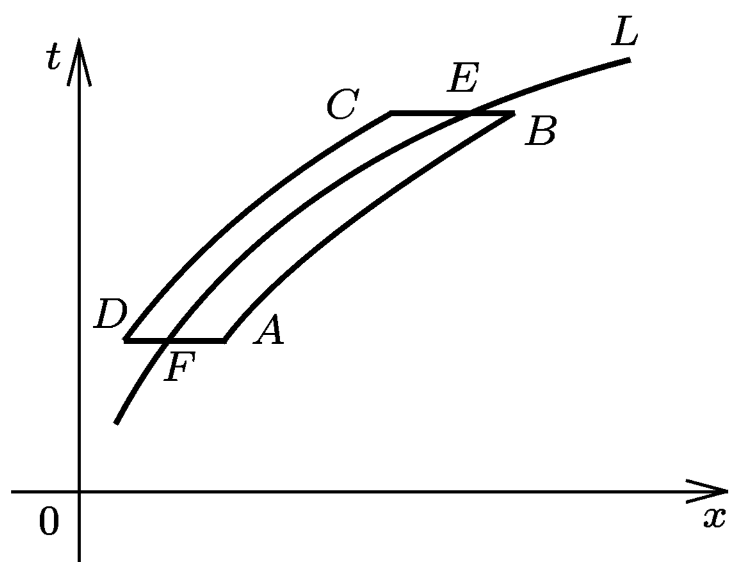


Рис. 32

ния точки разрыва $\dot{x} = dx/dt$ не могут быть произвольными: они связаны между собой некоторым соотношением.

Пусть L является линией разрыва (рис. 32). Интеграл

$$\int_\gamma \frac{u^k}{k}\,dx - \frac{u^{k+1}}{k+1}\,dt$$

по контуру $ABCDA$, как и по любому другому контуру, обращается в нуль. Когда отрезки BC и DA стягиваются к точкам E и F соответственно, интегралы по ним обращаются в нуль и получается равенство

$$\int_{L'} \left[\frac{u^k}{k}\right] dx - \left[\frac{u^{k+1}}{k+1}\right] dt = 0,$$

или

$$\int_{L'} \left(\left[\frac{u^k}{k}\right]\frac{dx}{dt} - \left[\frac{u^{k+1}}{k+1}\right]\right) dt = 0,$$

где $[z] = z_{\text{прав.}} - z_{\text{лев.}}$ — скачок величины z на линии разрыва, а L' — произвольный участок этой линии.

В силу произвольности участка L' в каждой точке должна обращаться в нуль подынтегральная функция:

$$\left[\frac{u^k}{k}\right]\frac{dx}{dt} - \left[\frac{u^{k+1}}{k+1}\right] = 0.$$

Тогда

$$\frac{dx}{dt} = \left[\frac{u^{k+1}}{k+1}\right] \Big/ \left[\frac{u^k}{k}\right]. \tag{6}$$

Отсюда при различных k на линии разрыва получаются различные условия. В частности, в случае $k = 1$

$$\frac{dx}{dt} = \frac{u_{\text{лев.}} + u_{\text{прав.}}}{2}, \tag{7}$$

а в случае $k = 2$

$$\frac{dx}{dt} = \frac{2}{3}\,\frac{u_{\text{лев.}}^2 + u_{\text{лев.}}u_{\text{прав.}} + u_{\text{прав.}}^2}{u_{\text{лев.}} + u_{\text{прав.}}}.$$

Отсюда видно, что условия на разрыве, которым должно удовлетворять разрывное решение задачи (А), (Б), зависит от k.

4. Обобщенное (разрывное) решение дифференциальной краевой задачи. Введем понятие обобщенного решения задачи Коши (1), отождествив это решение с решением задачи (А), (Б).

В случае решения, имеющего, как мы видели, всюду непрерывные производные, это решение не зависит от k и совпадает с обычным решением задачи Коши (1); решение в этом случае — это дифференцируемая функция, обращающая уравнение $\partial u/\partial t + \partial u/\partial x = 0$ в тождество и удовлетворяющая начальному условию $u(x, 0) = \psi(x)$. В плодотворности рассмотрения наряду с задачей (А), (Б) равносильной ей задачи Коши (1) мы убедились, занимаясь выяснением механизма возникновения разрывов. Этот механизм трудно было бы выяснить, оставаясь в рамках постановки задачи (А), (Б) и не используя задачу Коши (1).

В случае разрывного решения приведенное нами определение обобщенного решения задачи (1) пока ничем не обогащает постановку задачи (А), (Б), являясь лишь ее переименованием. Поэтому придадим определению обобщенного решения задачи (1), приведенному выше, другую форму.

При этом мы будем рассматривать лишь те решения задачи (А), (Б), которые в полосе $0 < t < T$ имеют непрерывные первые производные всюду, кроме некоторых гладких кривых $x = x(t)$, на которых они могут претерпевать разрывы первого рода (скачки).

Будем называть $u(x, t)$ *обобщенным решением задачи Коши* (1), *соответствующим интегральному закону сохранения* (А), если:

а) функция $u(x, t)$ удовлетворяет дифференциальному уравнению (1) в каждой точке полосы $0 < t < T$, не лежащей на линиях разрыва;

б) на линии разрыва выполняется условие (6);

в) для каждого x, для которого $\psi(x)$ непрерывна, функция $u(x, t)$ непрерывна в точке $(x, 0)$ и удовлетворяет начальному условию $u(x, 0) = \psi(x)$.

Мы не будем доказывать равносильность двух определений обобщенного решения, приведенных выше, предоставляя это читателю в качестве упражнения.

Подчеркнем, что обобщенное решение задачи Коши (1) задается в случае разрывного решения не только равенствами (1), но и указанием того, какому интегральному закону сохранения это обобщенное решение соответствует. В нашей постановке задачи (А), (Б) нет оснований предпочесть какое-либо значение k.

В задачах математической физики интегральный закон сохранения, аналогичный условию (А), выражает некоторое конкретное реальное свойство и вполне определен.

В дальнейших рассмотрениях мы будем считать для определенности, что выполняются интегральный закон сохранения (А), отвечаю-

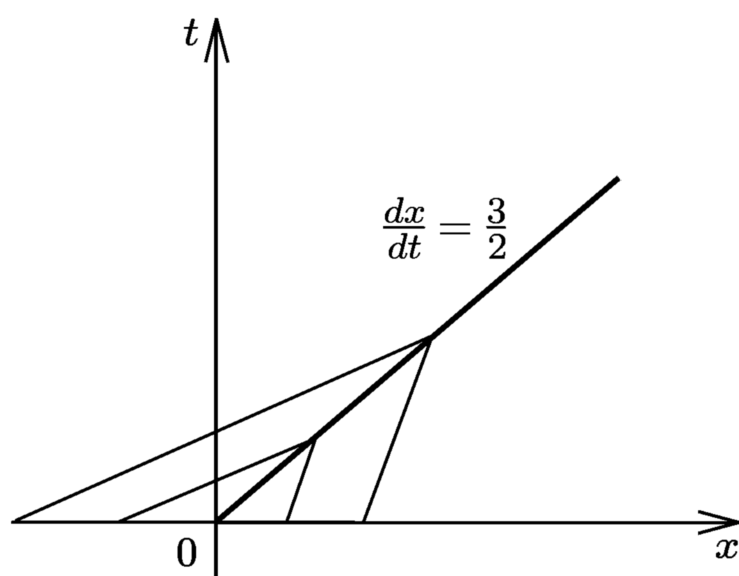


Рис. 33

щий значению $k = 1$, и вытекающее из него условие (7) на разрыве.

5. Распад произвольного разрыва. Пусть заданы разрывные начальные данные

$$u = \begin{cases} 2, & x < 0, \\ 1, & x > 0. \end{cases}$$

Построенное по этим начальным данным решение изображено на рис. 33. Тангенс угла наклона линии разрыва $dx/dt = (2+1)/2 = 3/2$ является средним арифметическим тангенсов углов наклона характеристик по обе стороны от нее.

Зададим теперь в начальных условиях другой разрыв:

$$u = \begin{cases} 1, & x < 0, \\ 2, & x > 0. \end{cases}$$

На рис. 34 видно, что возможны два способа построения решения.

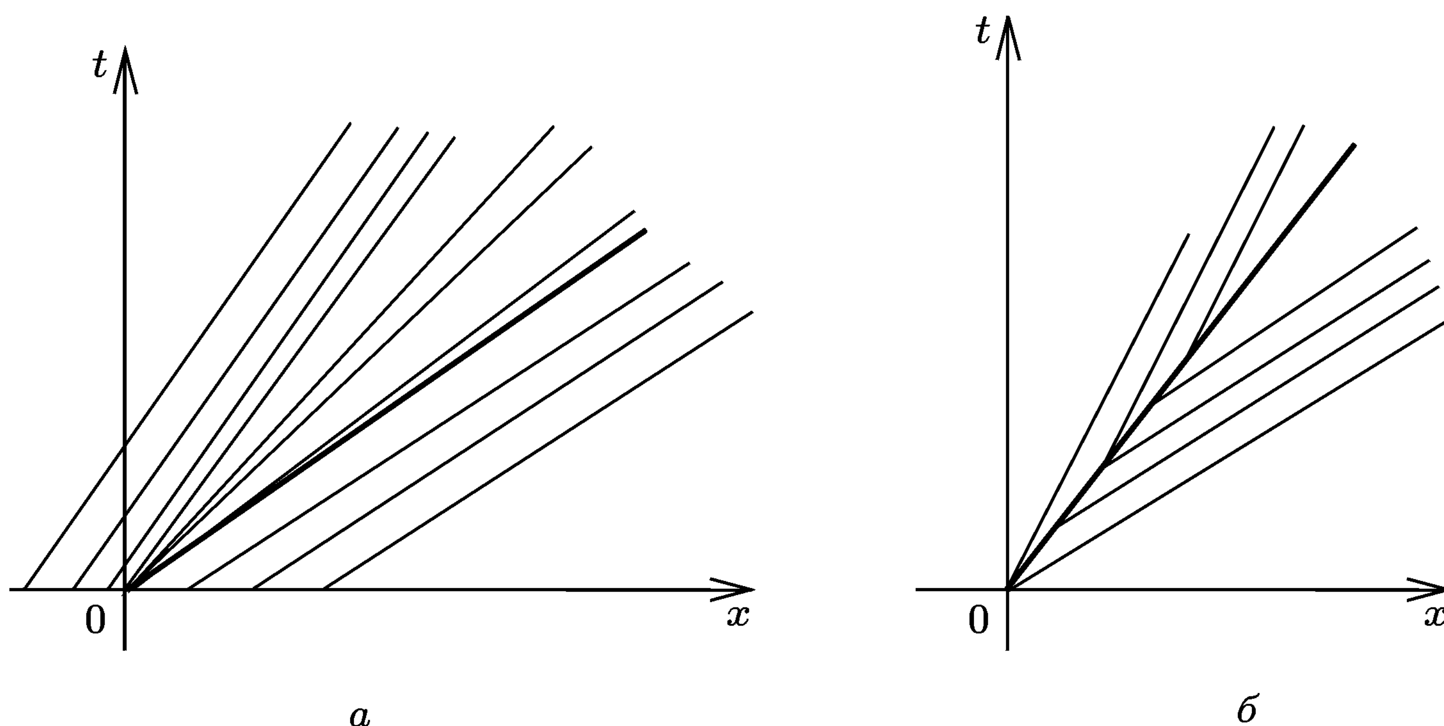


Рис. 34

При первом способе (рис. 34, *а*) мы получаем непрерывное решение, а при втором (рис. 34, *б*) — разрывное решение при $t > 0$. Следует предпочесть непрерывное решение. В пользу этого говорит следующее рассуждение. Если несколько изменить начальные данные, задав их формулой

$$u = \begin{cases} 1, & x \leqslant 0, \\ 2, & x \geqslant \varepsilon, \\ 1 + x/\varepsilon, & 0 \leqslant x \leqslant \varepsilon, \end{cases}$$

то решение u определится однозначно. Оно изображено на рис. 35. При стремлении ε к нулю это решение переходит в непрерывное ре-

шение, изображенное на рис. 34, *а*. Запрет решения, изображенного на рис. 34, *б*, по причине его неустойчивости относительно возмущения начальных данных аналогичен запрету ударных волн разрежения при математическом описании течения идеального газа.

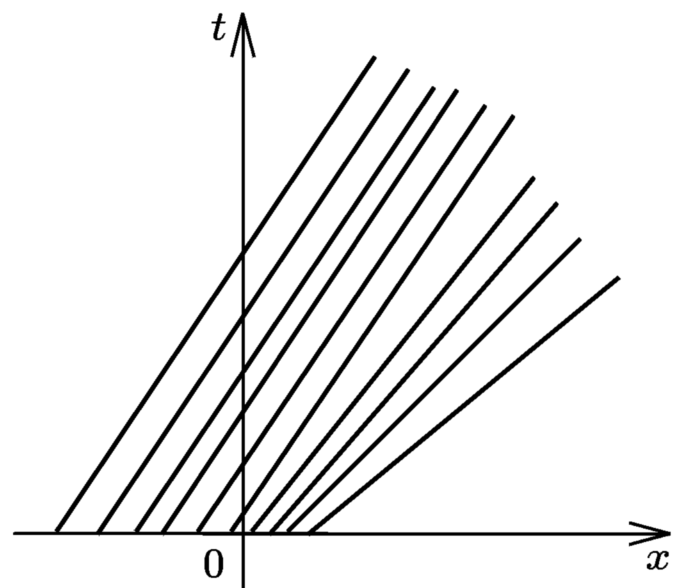


Рис. 35

Задача*

Рассмотрим вспомогательную задачу

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \mu\frac{\partial^2 u}{\partial x^2}, \quad u(x,0) = \psi(x), \quad -\infty < x < \infty, \tag{8}$$

где $\mu > 0$ — некоторый параметр, который аналогичен вязкости в случае задач газовой динамики. Уравнение (8) параболического типа и имеет гладкое решение при любой гладкой функции $\psi(x)$.

Доказать, что при $\mu \to 0$ в случае $\psi(x) = 2$ при $x < 0$ и $\psi(x) = 1$ при $x > 0$ решение стремится к обобщенному в смысле данного выше определения решению, отвечающему задаче (А), (Б) при $k = 1$.

§ 2. Построение разностных схем

Перейдем теперь к вопросу о построении разностных схем для вычисления обобщенного решения задачи

$$\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = 0, \quad u(x,0) = \psi(x), \tag{1}$$

соответствующего интегральному закону сохранения (3) из § 1 при $k = 1$.

Будем предполагать для определенности, что $\psi(x) > 0$. Тогда $u(x,t) > 0$. Первое, что кажется естественным, это рассмотреть разностную схему

$$\begin{gathered} \frac{u_m^{p+1} - u_m^p}{\tau} + u_m^p\frac{u_m^p - u_{m-1}^p}{h} = 0, \\ m = 0, \pm 1, \ldots, \quad p = 0, 1, \ldots, \\ u_m^0 = \psi(mh). \end{gathered} \tag{2}$$

Замораживая коэффициент u_m^p в точке $m = m_0$, мы видим, что для возникающего уравнения с постоянными коэффициентами при переходе на слой $t = (p+1)\tau$ выполняется принцип максимума, если шаг $\tau = \tau_p$ выбран из условия

$$r_p = \frac{\tau_p}{h} \leqslant \frac{1}{\max\limits_m |u_m^p|}.$$

Поэтому можно ожидать устойчивости. Если решение задачи (1) гладкое, то аппроксимация задачи (1) задачей (2) не вызывает сомнения. Действительно, в этом случае экспериментальные расчеты заранее известных гладких решений подтверждают сходимость.

Однако если задача (1) имеет разрывное решение, то сходимости к обобщенному решению ни в каком разумном смысле ожидать нет оснований. Ведь в используемую разностную схему (2) не заложена информация о том, какой именно закон сохранения положен в основу определения обобщенного решения. Здесь нарушено условие Куранта, Фридрихса и Леви в том смысле, что обобщенное решение зависит от числа k, определяющего интегральный закон сохранения (А), а решение разностной схемы не зависит.

Изложим два подхода к построению алгоритмов для вычисления обобщенного решения.

1. Метод характеристик. В этом методе развитие возникающих в процессе расчета, т. е. при увеличении времени t, разрывов ведется по особым формулам, основанным на условии (7) из § 1, которое должно выполняться на разрыве. В точках областей гладкости используется запись закона сохранения в дифференциальной форме, т. е. уравнение

$$\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} = 0.$$

Принципиальная схема одного из вариантов метода характеристик применительно к нашему примеру состоит в следующем. Отметим на оси Ox точки $x_m = mh$. Будем считать для определенности, что начальное условие $u(x, 0) = \psi(x)$ задается гладкой функцией $\psi(x)$. Из каждой точки $(x_m, 0)$ выпустим характеристику уравнения $u_t + u u_x = 0$.

Предположим, чтобы не осложнять изложение, что при заданной функции $\psi(x)$ можно выбрать столь малое τ, что на любом отрезке времени длины τ каждая характеристика пересекается не более чем с одной из соседних характеристик. Возьмем такое τ и проведем прямые $t = t_p = p\tau$. Рассмотрим точки пересечения характеристик, выходящих из точек $(x_m, 0)$, с прямой $t = \tau$ и перенесем в эти точки значения решения $u(x_m, 0) = \psi(x_m)$ по характеристикам.

Если на участке $0 \leqslant t \leqslant \tau$ никакие две характеристики не пересекались, то делаем следующий шаг: продолжаем характеристики до пересечения с прямой $t = 2\tau$ и переносим по характеристикам значения решения в точки пересечения. Если пересечения характеристик за время $\tau < t < 2\tau$ опять не было, то делаем следующий шаг, и так до тех пор, пока на некотором участке $t_p \leqslant t \leqslant t_{p+1}$ не пересекутся две характеристики (например, выходящие из точек $(x_m, 0)$ и $(x_{m+1}, 0)$)

(рис. 36). Тогда середину отрезка $Q_{m+1}^{p+1}Q_m^{p+1}$ будем считать точкой,

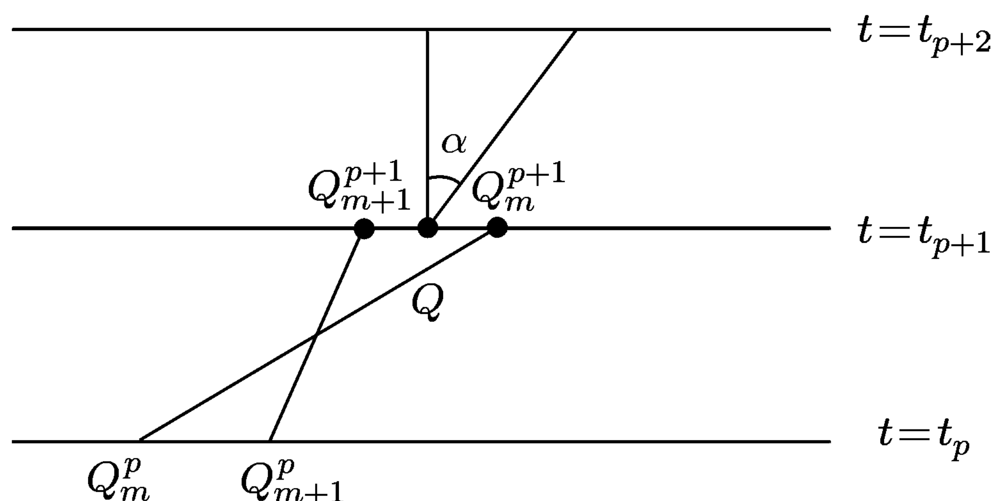


Рис. 36

из которой выходит зарождающийся разрыв. Точки Q_{m+1}^{p+1} и Q_m^p заменим одной точкой Q, приписывая ей два значения решения: $u_{\text{лев.}}$ и $u_{\text{прав.}}$, принимая за эти величины значения

$$u_{\text{лев.}} = u(Q_m^{p+1}), \quad u_{\text{прав.}} = u(Q_{m+1}^{p+1}).$$

Из точки Q выпускаем линию разрыва до пересечения с прямой $t = t_{p+2}$. Угловой коэффициент разрыва определяем из условия на разрыве

$$\operatorname{tg}\alpha = \frac{u_{\text{лев.}} + u_{\text{прав.}}}{2}.$$

Из точки пересечения разрыва с прямой $t = t_{p+2}$ проводим характеристики назад до пересечения с прямой $t = t_{p+1}$ с угловыми коэффициентами $u_{\text{лев.}}$, $u_{\text{прав.}}$ с предыдущего слоя. В точках пересечения этих характеристик с прямой $t = t_{p+1}$ с помощью интерполяции по x находим значения u и принимаем их за левое и правое значения решения в точке разрыва, лежащей на прямой $t = t_{p+2}$. Это позволяет определить новый наклон разрыва как среднее арифметическое найденных значений слева и справа и продолжить разрыв еще на шаг τ по времени.

Достоинство метода характеристик состоит в том, что он позволяет следить за разрывами и аккуратно их рассчитывать. Однако в процессе счета возникают все новые разрывы; в частности, малосущественные разрывы могут пересекаться, так что с течением времени картина усложняется. Логика расчета усложняется, требования к машинной памяти и расход машинного времени возрастают. В этом заключается недостаток метода характеристик, при котором разрывы выделены и считаются нестандартным образом.

2. Дивергентные разностные схемы. Разностные схемы, не использующие искусственно введенную вязкость и не использующие условия на разрыве, как было выяснено, должны опираться на интегральные законы сохранения.

Проведем на плоскости Oxt сетку прямых $t = p\tau$, $x = (m + 1/2)h$

$(p = 0, 1, \ldots, m = 0, \pm 1, \ldots)$. Отметим на сторонах возникающих пря-

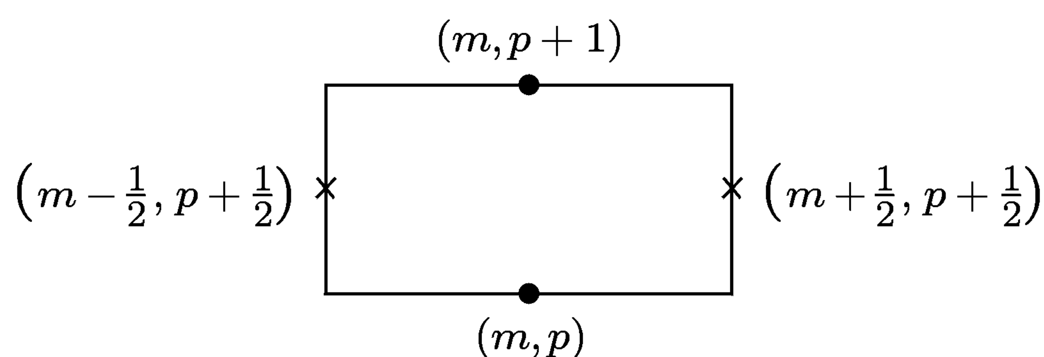


Рис. 37

моугольников их середины (рис. 37) и отнесем их к сетке D_h (оси координат на рисунке не показаны).

Искомой функцией $[u]_h$ будем считать сеточную функцию, определенную в точках сетки D_h путем усреднения решения $u(x,t)$ по той стороне сеточного прямоугольничка, которому принадлежит рассматриваемая точка сетки:

$$[u]_h\Big|_{\substack{x=x_m\\ t=t_p}} \equiv \widetilde{u}_m^p = \frac{1}{h}\int\limits_{x_{m-1/2}}^{x_{m+1/2}} u(x,t_p)\,dx,$$

$$[u]_h\Big|_{\substack{x=x_{m+1/2}\\ t=t_{p+1/2}}} \equiv \widetilde{U}_{m+1/2}^{p+1/2} = \frac{1}{\tau}\int\limits_{t_p}^{t_{p+1}} u(x_{m+1/2},t)\,dt.$$

Приближенное решение $u^{(h)}$ задачи определено на той же сетке D_h. Значения $u^{(h)}$ в точках (x_m, t_p), лежащих на горизонтальных сторонах прямоугольников, будем обозначать через u_m^p, а в точках $(x_{m+1/2}, t_{p+1/2})$ вертикальных сторон — через $U_{m+1/2}^{p+1/2}$.

Величину u_m^p можно считать продолженной на всю сторону $t = t_p$, $x_{m-1/2} < x < x_{m+1/2}$ прямоугольника, которому принадлежит точка (x_m, t_p). Аналогично будем считать, что $U_{m+1/2}^{p+1/2}$ определена на всем вертикальном промежутке

$$x = x_{m+1/2}, \quad t_p < t < t_{p+1}.$$

Таким образом, $u^{(h)}$ будет функцией, определенной на прямых $x = (m+1/2)h$, $t = p\tau$. Связь между величинами u_m^p и $U_{m+1/2}^{p+1/2}$ $(p = 0, 1, \ldots, m = 0, \pm 1, \ldots)$ установим, исходя из интегрального закона сохранения

$$\oint\limits_{\Gamma} u\,dx - \frac{u^2}{2}\,dt = 0.$$

Рассмотрим в качестве контура Γ элементарный прямоугольник сетки и положим

$$-\oint\limits_{\Gamma} u^{(h)} dx - \frac{|u^{(h)}|^2}{2}\,dt = 0, \tag{3}$$

или, в развернутом виде,

$$h\,[u_m^{p+1} - u_m^p] + \frac{\tau}{2}\,[(U_{m+1/2}^{p+1/2})^2 - (U_{m-1/2}^{p+1/2})^2]\,dt = 0. \qquad (4)$$

Если будет указано правило вычисления величин $U_{m+1/2}^{p+1/2}$ ($m = 0, \pm 1, \ldots$) по уже известным величинам u_m^p ($m = 0, \pm 1, \ldots$), то формула (4) позволит вычислить величины u_m^{p+1} ($m = 0, \pm 1, \ldots$), т. е. продвинуться на один шаг по времени. Однако независимо от конкретного способа, который мы изберем для вычисления величины $U_{m+1/2}^{p+1/2}$, разностная схема вида (4) обладает свойством дивергентности, которое состоит в следующем.

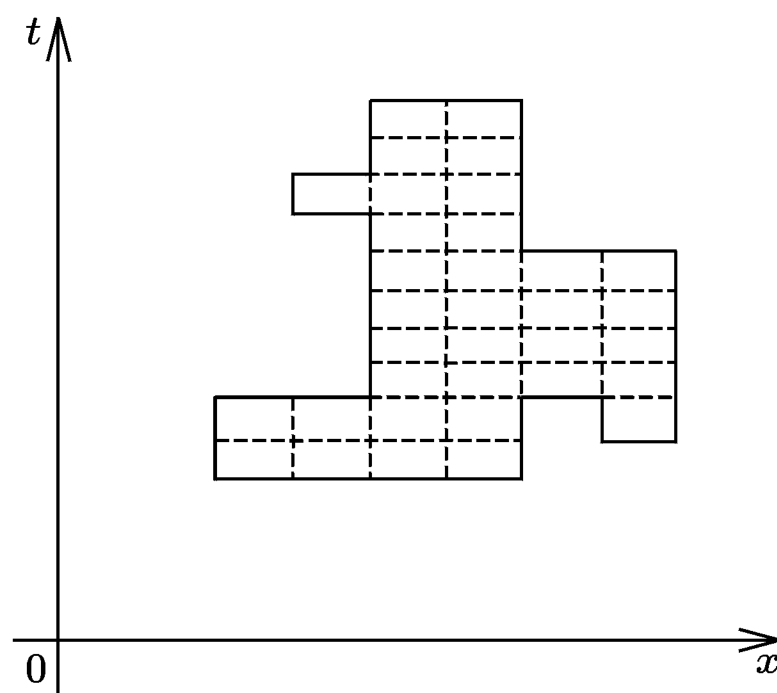
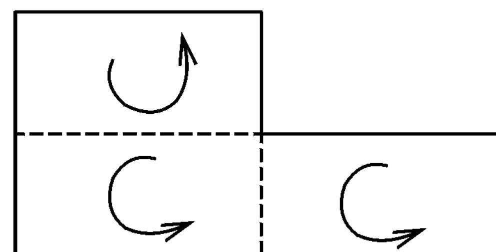


Рис. 38

Проведем в полуплоскости $t \geqslant 0$ какой-либо замкнутый несамопересекающийся контур, целиком состоящий из сторон сеточных прямоугольников (рис. 38). Этот контур g_h ограничит некоторую область G_h, составленную из сеточных прямоугольников. Просуммируем почленно все уравнения (4), относящиеся к прямоугольникам, составляющим область G_h. Уравнения (5) и (3) отличаются только формой записи. Поэтому можно считать, что мы суммируем уравнения (3). Получим

$$-\oint_{g_h} u^{(h)} dx - \frac{|u^{(h)}|^2}{2}\,dt = 0. \qquad (5)$$

Интегралы по тем сторонам прямоугольников, которые не лежат на границе g_h области G_h, но входят в выражение (3) после суммирования уравнений (3), взаимно уничтожаются. Действительно, каждая из этих сторон принадлежит двум соседним прямоугольникам, так что интегрирование функции $u^{(h)}$ по каждой из этих строк встречается дважды и ведется в противоположных направлениях (рис. 39).


Рис. 39

Разностные схемы, при суммировании которых по точкам сеточной области G_h остаются только алгебраические суммы значений неизвестных или функций от них вдоль границы области, называют *дивергентными* или *консервативными*. Такие схемы аналогичны дифференциальным уравнениям дивергентного вида

$$\operatorname{div}\,\Phi = \frac{\partial \Phi_1}{\partial t} + \frac{\partial \Phi_2}{\partial x} = 0,$$

при почленном интегрировании которых по двумерной области D в левой части возникает контурный интеграл (2) из § 1. Разностная схема (2) недивергентна, схема (4) дивергентна.

Заметим следующее. Пусть сеточная функция $u^{(h)}$, удовлетворяющая уравнению (4), при $h \to 0$ равномерно сходится к некоторой кусочно непрерывной функции $u(x,t)$ во всякой замкнутой области, не содержащей линий разрыва, и пусть $u^{(h)}$ равномерно по h ограничена. Тогда $u(x,t)$ удовлетворяет интегральному закону сохранения

$$\oint_g u\,dx - \frac{u^2}{2}\,dt = 0,$$

где g — произвольный кусочно гладкий контур, т. е. $u^{(h)}$ сходится к обобщенному решению. Это непосредственно следует из возможности приблизить контур g контуром g_h, из равенства (5) и предположенной сходимости *).

Чтобы схема (4) приобрела смысл, надо указать способ вычисления величин $U^{p+1/2}_{m+1/2}$ по величинам u^p_m. В схеме Годунова, которую мы взяли для иллюстрации понятия дивергентных схем, для этого используется решение следующей задачи о распаде разрыва.

Пусть в начальный момент решение $u(x,0)$ задано условиями

$$u(x,0) = \begin{cases} u_{\text{лев.}}, & x < 0, \\ u_{\text{прав.}}, & x > 0, \end{cases}$$

где $u_{\text{лев.}} = \text{const}$, $u_{\text{прав.}} = \text{const}$. Тогда можно найти соответствующее обобщенное решение. Как это делается, мы видели в § 1 при разборе примера $u_{\text{лев.}} = 1$, $u_{\text{прав.}} = 2$ и примера $u_{\text{лев.}} = 2$, $u_{\text{прав.}} = 1$. Нам важно знать значение $U = u(0,t)$ решения $u(x,t)$ при $x = 0$.

Читатель, построив картинки типа рис. 28, изображающие решение $u(x,t)$, легко проверит, что на прямой $x = 0$ решение принимает значения $u_{\text{лев.}}$, $u_{\text{прав.}}$ или 0 в зависимости от заданных начальных данных, и выяснит для каждой конкретной пары чисел $u_{\text{лев.}}$, $u_{\text{прав.}}$, какое именно. Например, при $u_{\text{лев.}} > 0$, $u_{\text{прав.}} > 0$ будет $u(0,t) = u_{\text{лев.}}$, а при $u_{\text{лев.}} < 0$, $u_{\text{прав.}} < 0$ будет $u(0,t) = u_{\text{прав.}}$.

Величину $U^{p+1/2}_{m+1/2} (= U)$ в схеме (4) будем определять из задачи о распаде разрыва, возникающего на границе $x = x_{m+1/2}$ каждых двух участков, где заданы постоянные значения $u^p_m (= u_{\text{лев.}})$, $u^p_{m+1} (= u_{\text{прав.}})$.

Если, например, $u^p_m > 0$ $(m = 0, \pm 1, \ldots)$, то $U^{p+1/2}_{m+1/2} = u_{\text{лев.}} = u^p_m$

*) Функция $u = u(x,t)$ определена почти всюду, а функция $u^{(h)}(x,t)$ — лишь на сетке прямых. Это формальное несоответствие можно преодолеть, считая, что при уменьшении h каждая новая сетка является подразделением старой, и говоря о сходимости в точках сетки, построенной для любого фиксированного h из числа допустимых.

($m = 0, \pm 1, \ldots$), и схема (4) примет вид

$$\frac{u_m^{p+1} - u_m^p}{\tau} + \frac{1}{h}\left[\frac{(u_m^p)^2}{2} - \frac{(u_{m-1}^p)^2}{2}\right] = 0,$$

$$u_m^0 = \frac{1}{h} \int\limits_{x_{m-1/2}}^{x_{m+1/2}} \psi(x)\, dx,$$

или

$$\frac{u_m^{p+1} - u_m^p}{\tau} + \frac{u_{m-1}^p + u_m^p}{2} \, \frac{u_m^p - u_{m-1}^p}{h} = 0.$$

Легко видеть, что при

$$r \leqslant \frac{\tau}{h} \leqslant \frac{1}{\max\limits_m |u_m^p|}$$

имеет место принцип максимума

$$\max_m |u_m^{p+1}| \leqslant \max_m |u_m^p| \leqslant \ldots \leqslant \max_m |u_m^0| \leqslant \max_x |\psi(x)|.$$

Отсюда видно, что при $\tau \leqslant \dfrac{1}{\max\limits_x |\psi(x)|}\, h$ можно надеяться, что полученная разностная схема устойчива при некотором разумном выборе норм. Мы не будем фактически указывать эти нормы: экспериментальные расчеты подтверждают, что при измельчении сетки решение $u^{(h)}$ задачи (4) с кусочно монотонными и кусочно гладкими начальными данными $\psi(x)$ сходится к некоторой функции $u(x, t)$, имеющей конечное число разрывов, причем вне любой окрестности разрывов сходимость равномерная.

Схема (4) с вычислением $U_{m+1/2}^{p+1/2}$ путем использования распада разрыва не является, конечно, единственной дивергентной схемой для задачи (1); см. [4, 9, 11].

ГЛАВА 11

РАЗНОСТНЫЕ МЕТОДЫ ДЛЯ ЭЛЛИПТИЧЕСКИХ ЗАДАЧ

Простейшим эллиптическим уравнением является уравнение Пуассона

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \varphi(x, y). \tag{1}$$

Пусть задана некоторая область D на плоскости, а на ее границе $\Gamma = \partial D$ поставлено краевое условие вида

$$\left(au - b\,\frac{\partial u}{\partial n}\right)\Big|_\Gamma = \psi(s), \tag{2}$$

где $\partial/\partial n$ — производная в направлении внутренней нормали, $a \geqslant 0$, $b \geqslant 0$ $(a^2 + b^2 = 1)$ — некоторые числа, s — длина дуги, отсчитываемая вдоль границы Г. Функции $\varphi(x, y)$, $\psi(s)$ считаются заданными. Требуется вычислить решение (1), (2) краевой задачи (1), (2). В случае $a = 1$, $b = 0$ возникающая задача называется *первой краевой задачей* или *задачей Дирихле.* В случае $a = 0$, $b = 1$ — это *вторая краевая задача*, или *задача Неймана.* В случае $a > 0$, $b > 0$ — это *третья краевая задача.*

Перечисленные три краевые задачи являются основными для уравнения Пуассона, но в математической физике возникают также и другие задачи для уравнения Пуассона. Наряду с уравнением Пуассона могут возникать уравнения с переменными коэффициентами вида

$$\frac{\partial}{\partial x}\left(a\,\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(b\,\frac{\partial u}{\partial y}\right) = \varphi(x, y), \tag{3}$$

где $a = a(x, y) > 0$, $b = b(x, y) > 0$, для которых также корректно поставлены задачи с краевыми условиями вида (2).

Эллиптические уравнения и системы эллиптических уравнений возникают при математическом моделировании многих стационарных состояний. Так, например, уравнение Пуассона может описывать потенциал электрического поля, потенциал скоростей установившегося потока несжимаемой жидкости, установившуюся температуру в однородном теплопроводном теле. Система уравнений упругости Ламе описывает смещения в находящемся под действием стационарных сил теле и т. д.

Численное решение краевых задач для эллиптических уравнений осуществляется во многих случаях с помощью разностных схем.

В простейших случаях, когда решение во всей рассматриваемой области меняется более или менее равномерно, а сама область не имеет узких перешейков, можно пользоваться регулярными сетками, а для получения разностных схем заменять производные разностными отношениями.

В противном случае регулярная сетка становится практически непригодной, так как места быстрого изменения решения или места, где область D имеет узкие горловины, диктуют слишком мелкую сетку. Но в случае нерегулярной сетки построение разностной схемы путем замены производных разностными отношениями становится невозможным. Для построения разностных схем на нерегулярных сетках большое распространение получили так называемые вариационно-разностные и проекционно-разностные схемы. Среди прикладников соответствующие схемы называют обычно *методом конечных элементов.*

В этой главе мы остановимся на трех вопросах. Во-первых, проверим, что простейшая пятиточечная разностная схема для задачи Дирихле в прямоугольной области, с которой мы уже не раз встречались,

обладает свойствами аппроксимации и устойчивости, а значит, является сходящейся при измельчении сетки. Во-вторых, дадим представление о методе конечных элементов. Наконец, объясним, как можно вычислять решения больших систем линейных уравнений, которыми являются при мелких сетках как простейшие разностные схемы, так и системы, возникающие в методе конечных элементов.

При этом в дополнение к изложенным в части II точным и итерационным методам решения систем линейных уравнений будет изложен многосеточный метод Федоренко.

§ 1. Аппроксимация и устойчивость простейшей разностной схемы

Рассмотрим задачу Дирихле для уравнения Пуассона в квадратной области $\overline{D} = (0 \leqslant x,\ y \leqslant 1)$ с границей Γ:

$$\begin{gathered} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \varphi(x,y), \quad 0 \leqslant x, \quad y \leqslant 1, \\ u\big|_{\Gamma} = \varphi(s). \end{gathered} \tag{1}$$

Совокупность точек $(x_m, y_n) = (mh, nh)$ сетки (где $h = 1/M$, M целое), попавших внутрь квадрата или на его границу, обозначим D_h.

Точки D_h, лежащие строго внутри квадрата D, будем называть *внутренними точками* сеточного квадрата D_h; совокупность внутренних точек обозначим D_h^0. Совокупность точек D_h, попавших на границу квадрата $\overline{D}$, будем обозначать Γ_h.

Рассмотрим разностную схему

$$L_h u^{(h)} = f^{(h)}; \tag{2}$$

здесь

$$L_h u^{(h)} = \begin{cases} \dfrac{u_{m+1,n} - 2u_{mn} + u_{m-1,n}}{h^2} + \dfrac{u_{m,n+1} - 2u_{mn} + u_{m,n-1}}{h^2} = \\ \qquad\qquad\qquad\qquad\qquad\qquad = \varphi(x_m, y_n), \\ u_{mn} = \psi(s_{mn}), \quad (x_m, y_n) \in \Gamma_h, \end{cases} \tag{3}$$

где $\psi(s_{mn})$ — значение функции $\psi(s)$ в точке (x_m, y_n), принадлежащей границе Γ_h.

1. Аппроксимация. Правая часть $f^{(h)}$ разностной схемы (2) имеет вид

$$f^{(h)} = \begin{cases} \varphi(x_m, y_n), \ (x_m, y_n) \in D_h^0, \\ \psi(s_{mn}), \quad\ (x_m, y_n) \in \Gamma_h. \end{cases} \tag{4}$$

В предположении, что решение $u(x,y)$ задачи (1) имеет ограниченные четвертые производные, с помощью формулы Тейлора устанавливается равенство

$$\frac{u(x+h,y)-2u(x,y)+u(x-h,y)}{h^2}+\frac{u(x,y+h)-2u(x,y)+u(x,y-h)}{h^2}=$$
$$=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{h^2}{24}\left(\frac{\partial^4 u(x+\xi,y)}{\partial x^4}+\frac{\partial^4 u(x,y+\eta)}{\partial y^4}\right). \quad (5)$$

Поэтому для решения $u(x,y)$ задачи (1) имеем

$$L_h[u]_h=\begin{cases}\varphi(x_m,y_n)+O(h^2), & (x_m,y_n)\in D_h^0,\\ \psi(s_{mn})+0, & (x_m,y_n)\in\Gamma_h.\end{cases} \quad (6)$$

Таким образом, невязка $\delta f^{(h)}$, возникающая при подстановке $[u]_h$ в левую часть разностной схемы (2), имеет вид

$$\delta f^{(h)}=\begin{cases}O(h^2), & (x_m,y_n)\in D_h^0,\\ 0, & (x_m,y_n)\in\Gamma_h.\end{cases} \quad (7)$$

В пространстве F_h, состоящем из элементов вида

$$f^{(h)}=\begin{cases}\varphi_{mn}, & (mh,nh)\in D_h^0,\\ \psi_{mn}, & (mh,nh)\in\Gamma_h,\end{cases}$$

введем норму

$$\|f^{(h)}\|_{F_h}=\max_{(mh,nh)\in D_h^0}|\varphi_{mn}|+\max_{(mh,nh)\in\Gamma_h}|\psi_{mn}|. \quad (8)$$

Тогда

$$\|\delta f^{(h)}\|_{F_h}=O(h^2).$$

Таким образом, разностная краевая задача (3) аппроксимирует задачу Дирихле (1) со вторым порядком относительно h.

2. Устойчивость. Определим норму в пространстве U_h функций, заданных на сетке D_h, положив

$$\|u^{(h)}\|_{U_h}=\max_{(mh,nh)\in D_h}|u_{mn}|. \quad (9)$$

Для доказательства устойчивости разностной схемы (3), к которому мы переходим, в соответствии с определением устойчивости надо установить, что задача (2) однозначно разрешима при произвольной правой части $f^{(h)}$ (это свойство не зависит от выбора норм) и что выполнена оценка вида

$$\|u^{(h)}\|_{U_h}\leqslant c\,\|f^{(h)}\|_{F_h}, \quad (10)$$

где c не зависит ни от h, ни от $f^{(h)}$.

Лемма 1. *Пусть функция $v^{(h)} = \{v_{mn}\}$ определена на сетке D_h и во внутренних точках $(x_m, y_n) = (mh, nh) \in D_h^0$ удовлетворяет условию*

$$\Lambda_h v^{(h)}\big|_{(mh,nh)} \geqslant 0, \quad (mh, nh) \in D_h^0, \tag{11}$$

$$\Lambda_h v^{(h)}\big|_{(mh,nh)} \equiv \frac{v_{m+1,n} - 2v_{mn} + v_{m-1,n}}{h^2} + \frac{v_{m,n+1} - 2v_{mn} + v_{m,n-1}}{h^2}.$$

Тогда наибольшее на сетке D_h значение $v^{(h)}$ достигается хотя бы в одной точке границы Γ_h.

Доказательство. Допустим противное. Выберем среди точек сетки D_h, в которых $v^{(h)}$ достигает своего наибольшего значения, какую-нибудь одну точку (x_m, y_n), имеющую самую большую абсциссу. По нашему предположению, (x_m, y_n) — внутренняя точка, причем v_{mn} строго больше, чем $v_{m+1,n}$. В точке (mh, nh)

$$\Lambda_h v^{(h)}\big|_{(mh,nh)} \equiv$$
$$\equiv \frac{(v_{m+1,n} - v_{mn}) + (v_{m-1,n} - v_{mn}) + (v_{m,n+1} - v_{mn}) + (v_{m,n-1} - v_{mn})}{h^2} < 0,$$

поскольку выражение, стоящее в первых скобках в числителе, отрицательно, а выражения в остальных скобках неположительны. Получаем противоречие с (11). □

Лемма 2. *Пусть функция $v^{(h)} = \{v_{mn}\}$ определена на сетке D_h и во внутренних точках удовлетворяет условию*

$$\Lambda_h v^{(h)}\big|_{(mh,nh)} \leqslant 0, \quad (mh, nh) \in D_h^0. \tag{12}$$

Тогда наименьшее на сетке D_h значение $v^{(h)}$ достигается хотя бы в одной точке границы.

Лемма 2 доказывается аналогично лемме 1.

Теорема 1 (принцип максимума). *Каждое решение разностного уравнения*

$$\Lambda_h v^{(h)}\big|_{(mh,nh)} = 0, \quad (mh, nh) \in D_h^0, \tag{13}$$

достигает своих наибольшего и наименьшего значений в некоторых точках границы Γ_h.

Доказательство получается объединением утверждений лемм 1, 2.

Свойство решений разностного уравнения (13) аналогично свойству решений $v(x, y)$ уравнения Лапласа $v_{xx} + v_{yy} = 0$ принимать наибольшее и наименьшее значения на границе области, где эти решения определены.

Из принципа максимума следует, что задача

$$L_h u^{(h)} = \begin{cases} \Lambda_h u^{(h)}|_{(mh,nh)} = 0, & (mh, nh) \in D_h^0, \\ u^{(h)}|_{(mh,nh)} = 0, & (mh, nh) \in \Gamma_h, \end{cases} \tag{14}$$

имеет только нулевое решение $u^{(h)} = 0$, поскольку наибольшее и наименьшее значения этого решения получаются в точках границы Γ_h, где $u_{mn} = 0$. Следовательно, разностная краевая задача (2) однозначно разрешима при произвольной правой части.

Переходим к доказательству оценки (10). В силу формулы (5) для произвольного многочлена второй (и даже третьей) степени

$$P(x, y) = ax^2 + bxy + cy^2 + dx + ey + f$$

выполнено равенство

$$\Lambda_h P = \partial^2 P/\partial x^2 + \partial^2 P/\partial y^2, \tag{15}$$

так как четвертые производные от $P(x, y)$, входящие в выражение остаточного члена формулы (5), обращаются в нуль.

Используя функции φ_{mn}, ψ_{mn} из правой части системы (3) и фиксируя $R = \sqrt{2}$, строим вспомогательную функцию

$$P^{(h)}(x, y) = \frac{1}{4}\left[R^2 - (x^2 + y^2)\right] \max_{(mh,nh)\in D_h^0} |\varphi_{mn}| + \max_{(mh,nh)\in\Gamma_h} |\psi_{mn}|,$$

которую будем рассматривать только в точках сетки D_h. Это отражено значком h в обозначении $P^{(h)}(x, y)$. В силу (15)

$$\Lambda_h P^{(h)}\Big|_{\substack{x=mh\\ y=nh}} = -\max_{(rh,sh)\in D_h^0} |\varphi_{rs}|, \quad (mh, nh) \in D_h^0.$$

Поэтому разность решения $u^{(h)}$ задачи (3) и функции $P^{(h)}$ удовлетворяет в точках D_h^0 равенствам

$$\Lambda_h[u^{(h)} - P^{(h)}] = \Lambda_h u^{(h)} - \Lambda_h P^{(h)} = \varphi_{mn} + \max|\varphi_{mn}| > 0.$$

В силу леммы 1 разность $u^{(h)} - P^{(h)}$ принимает свое наибольшее значение на границе Γ_h. Но на границе Γ_h эта разность

$$u^{(h)}\Big|_{\Gamma_h} - P^{(h)}\Big|_{\Gamma_h} = \psi_{mn} - P^{(h)}(mh, nh) =$$
$$= \left[\psi_{mn} - \max_{(rh,sh)\in\Gamma_h} |\psi_{rs}|\right] + \frac{1}{4}\left|x^2 + y^2 - R^2\right| \max_{(rh,sh)\in D_h^0} |\varphi_{rs}|$$

неположительна, так как в квадрате D всюду $x^2 + y^2 \leqslant R^2$, и обе квадратные скобки в правой части неположительны. Поскольку наибольшее значение $u^{(h)} - P^{(h)}$ неположительно, то всюду на D_h^0

$$[u^{(h)} - P^{(h)}]_{mh,nh} \leqslant 0,$$

или

$$u^{(h)} \leqslant P^{(h)}.$$

Аналогично для функции $u^{(h)} + P^{(h)}$ в точках D_h^0

$$\Lambda_h(u^{(h)} + P^{(h)}) \leqslant 0,$$

а в точках Γ_h сумма $u^{(h)} + P^{(h)}$ неотрицательна. В силу леммы 2 всюду на D_h^0

$$u^{(h)} + P^{(h)} \geqslant 0,$$

или

$$-P^{(h)} \leqslant u^{(h)}.$$

Таким образом, всюду на D_h

$$|u_{mn}| \leqslant |P_{mn}^{(h)}| \leqslant \frac{1}{4} R^2 \max_{(rh,sh)\in D_h^0} |\varphi_{rs}| + \max_{(rh,sh)\in\Gamma_h} |\psi_{rs}|.$$

Отсюда вытекает неравенство (10):

$$\max |u_{mn}| = \|u^{(h)}\|_{U_h} \leqslant c\,\|f^{(h)}\|_{F_h} = c\left(\max_{(rh,sh)\in D_h^0} |\varphi_{rs}| + \max_{(rh,sh)\in\Gamma_h} |\psi_{rs}|\right),$$

где

$$c = \max\,[1, R^2/4],$$

завершающее доказательство устойчивости.

В случае задачи Дирихле для эллиптического уравнения с переменными коэффициентами

$$\frac{\partial}{\partial x}\left[a\,\frac{\partial u}{\partial x}\right] + \frac{\partial}{\partial y}\left[b\,\frac{\partial u}{\partial y}\right] = \varphi(x,y), \quad (x,y)\in D,$$

$$u\big|_\Gamma = \psi(s),$$

где $a = a(x,y)$, $b = b(x,y)$ — положительные в прямоугольнике D гладкие функции, разностную схему можно построить аналогично. Используя во внутренних точках сетки замену выражений $\dfrac{\partial}{\partial x}\left(a\,\dfrac{\partial u}{\partial x}\right)$, $\dfrac{\partial}{\partial y}\left(b\,\dfrac{\partial u}{\partial y}\right)$ разностными отношениями по приближенным формулам

$$\frac{\partial}{\partial x}\left[a(x,y)\,\frac{\partial u(x,y)}{\partial x}\right] \approx \widetilde{\Lambda}_{xx} u(x,y) =$$
$$= \frac{1}{h}\left[a\left(x+\frac{h}{2},y\right)\frac{u(x+h,y)-u(x,y)}{h} - a\left(x-\frac{h}{2},y\right)\frac{u(x,y)-u(x-h,y)}{h}\right],$$

$$\frac{\partial}{\partial y}\left[b(x,y)\,\frac{\partial u(x,y)}{\partial y}\right] \approx \widetilde{\Lambda}_{yy} u(x,y) =$$
$$= \frac{1}{h}\left[b\left(x,y+\frac{h}{2}\right)\frac{u(x,y+h)-u(x,y)}{h} - b\left(x,y-\frac{h}{2}\right)\frac{u(x,y)-u(x,y-h)}{h}\right],$$

получаем разностную схему (2) вида

$$L_h u^{(h)} = \begin{cases} \widetilde{\Lambda}_{xx} u^{(h)} + \widetilde{\Lambda}_{yy} u^{(h)} = \varphi(mh, nh), & (mh, nh) \in D_h^0, \\ u\big|_{\Gamma_h} = \psi(s_{mn}), \quad (mh, nh) \in \Gamma_h. \end{cases} \tag{16}$$

Пользуясь формулой Тейлора, можно убедиться в том, что имеется второй порядок аппроксимации. Можно было бы доказать устойчивость построенной схемы, преодолевая некоторые дополнительные трудности по сравнению с рассмотренными нами при разборе примера.

На практике при решении конкретных задач обычно ограничиваются обоснованиями принципиального характера на модельных задачах типа приведенных выше. Конкретные суждения о погрешности получаются, как правило, не из теоретических оценок, а из сравнения между собой результатов расчетов, выполненных на сетках с различными значениями шага h.

После того как разностная краевая задача, аппроксимирующая дифференциальную, построена, нужно еще указать не слишком трудоемкий способ ее решения. Ведь при малом h задача (16) есть система скалярных уравнений очень высокого порядка. В разобранном нами примере решение разностных уравнений — сложная и интересная задача, но мы отложим ее рассмотрение до § 3.

Задачи

1. Доказать, что если во внутренних точках области D_h функция $u^{(h)}$ удовлетворяет уравнению

$$\Lambda_h u^{(h)}\Big|_{(mh,nh)} = 0, \quad (mh, nh) \in D_h^0,$$

то либо $u^{(h)}$ принимает всюду на D_h одинаковые значения, либо наибольшее и наименьшее значения функции $u^{(h)}$ не достигаются ни в одной внутренней точке сетки D_h (*усиленный принцип максимума*).

2. Доказать, что если во всех внутренних точках области D_h выполнено условие $\Lambda_h u^{(h)} \geqslant 0$, причем хотя бы в одной точке неравенство строгое, то $u^{(h)}$ не достигает своего наибольшего значения ни в одной внутренней точке.

3. Рассмотрим разностную схему $L_h u^{(h)} = f^{(h)}$ вида

$$L_h u^{(h)} = \begin{cases} \Lambda_h u^{(h)}|_{(mh,nh)} = \varphi(mh, nh), & (mh, nh) \in D_h^0, \\ u_{mn} = \psi_1(s_{mn}), & m = M \text{ или } n = 0, \text{ или } n = M, \\ \dfrac{u_{1,n} - u_{0,n}}{h} = \psi_2(s_{mn}), & n = 1, 2, ..., M-1. \end{cases}$$

Эта разностная схема аппроксимирует задачу

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \varphi(x, y), \quad (x, y) \in D,$$

$$u(x, y) = \psi_1(s), \quad x = 1, \text{ или } y = 0, \text{ или } y = 1,$$

$$\frac{\partial u}{\partial x} = \psi_2(s), \quad x = 0.$$

а) Доказать, что при любых $\varphi(mh, nh) = 0$, $\psi_1(s_{mn})$, $\psi_2(s_{mn})$ задача $L_h u^{(h)} = f^{(h)}$ имеет единственное решение.

б*) Доказать, что если $\varphi(mh, nh)$ неотрицательно, а $\psi_1(s_{mn})$, $\psi_2(s_{mn})$ неположительны, то $u^{(h)}$ неположительно.

в*) Доказать, что при любых φ, ψ_1, ψ_2 имеет место оценка вида

$$\max_{(mh,nh)\in D_h} |u_{mn}| \leqslant c \left(\max_{(mh,nh)\in D_h^0} |\varphi_{mn}| + \max_{m,n} |\psi_1(s_{mn})| + \max_n |\psi_2(s_{0n})| \right),$$

где c — некоторая постоянная, не зависящая от h. Вычислить c.

§ 2. Понятие о методе конечных элементов

В этом параграфе мы изложим способ построения разностных схем, основанный на использовании той или иной вариационной постановки краевой задачи, решение которой требуется найти численно. Этот способ, называемый иногда *методом конечных элементов*, позволяет строить пригодные разностные схемы на нерегулярных сетках, а также при меньших предположениях о гладкости искомого решения и коэффициентов уравнения. Благодаря появляющейся свободе в выборе сеток узлы можно располагать гуще в тех частях области определения искомого решения, где решение ведет себя более сложно или где нас интересуют более мелкие детали его поведения. Возможность целесообразно располагать узлы позволяет достигать требуемой точности при меньшем числе узлов сетки.

Метод конечных элементов можно интерпретировать как одну из возможных конкретизаций классических вариационных методов решения краевых задач. Поэтому мы сначала расскажем о классическом вариационном методе, а затем о вариационно-разностных схемах.

1. Вариационная постановка краевых задач. Многие дифференциальные краевые задачи математической физики допускают естественные вариационные постановки. Мы ограничимся рассмотрением вариационной постановки лишь для одного примера, иллюстрирующего, однако, суть дела. Именно, речь пойдет о задаче Дирихле для уравнения Пуассона в некоторой ограниченной области D плоскости Oxy с кусочно гладкой границей Γ.

Обозначим через W линейное пространство всех непрерывных в области D и на ее границе Γ функций, обладающих ограниченными производными первого порядка, которые могут иметь разрывы лишь на конечном множестве прямых (для каждой функции $w(x, y) \in W$ своих). Введем в линейном пространстве W норму, положив для каждой функции $w \in W$

$$\|w\|_W = \left\{ \iint\limits_D w^2\,dx\,dy + \iint\limits_D \left[\left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 \right] dx\,dy \right\}^{1/2}. \qquad (1)$$

Пополнение пространства W приводит к полному пространству W_2^1 Соболева.

Рассмотрим первую краевую задачу (задачу Дирихле)

$$\begin{gathered}\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=f(x,y), \quad (x,y)\in D,\\ u\big|_\Gamma=\varphi(s),\end{gathered} \tag{A}$$

где s — длина дуги вдоль границы Γ области D, а $f(x,y)$, $\varphi(s)$ — заданные функции, удовлетворяющие всем условиям решения $u(x,y)$ задачи (A), обеспечивающим существование непрерывных вторых производных всюду в области D и на ее границе Γ.

Теорема 1. *Среди всех функций $w\in W$, удовлетворяющих граничному условию*

$$w\big|_\Gamma=\varphi(s), \tag{2}$$

решение $u(x,y)$ задачи (A) *придает выражению* (*функционалу*)

$$I(w)=\iint\limits_D\Big[\Big(\frac{\partial w}{\partial x}\Big)^2+\Big(\frac{\partial w}{\partial y}\Big)^2+2fw\Big]dx\,dy \tag{3}$$

наименьшее численное значение.

Доказательство. Пусть $w(x,y)\in w$ $(W\big|_\Gamma=\varphi(s))$ — некоторая фиксированная функция. Введем обозначение

$$\xi(x,y)\equiv w(x,y)-u(x,y).$$

Поскольку $u(x,y)$ имеет непрерывные вторые производные, а $w(x,y)\in\in W$, то также $\xi(x,y)\in W$, причем $\xi\big|_\Gamma=0$. Докажем равенство

$$I(w)=I(u+\xi)=I(u)+\iint\limits_D\Big[\Big(\frac{\partial \xi}{\partial x}\Big)^2+\Big(\frac{\partial \xi}{\partial y}\Big)^2\Big]dx\,dy, \tag{4}$$

из которого следует справедливость теоремы, поскольку в случае $w(x,y)\neq u(x,y)$ функция $\xi(x,y)$ не обращается тождественно в нуль, так что второе слагаемое в правой части формулы (4) строго положительно, и $I(w)>I(u)$. Очевидно, что

$$\begin{aligned}I(u+\xi)\equiv\iint\limits_D\Big[\Big(\frac{\partial u}{\partial x}+\frac{\partial \xi}{\partial x}\Big)^2+\Big(\frac{\partial u}{\partial y}+\frac{\partial \xi}{\partial y}\Big)^2+2(u+\xi)f\Big]dx\,dy=\\ =I(u)+\iint\limits_D\Big[\Big(\frac{\partial \xi}{\partial x}\Big)^2+\Big(\frac{\partial \xi}{\partial y}\Big)^2\Big]dx\,dy+\\ +2\iint\limits_D\Big[\frac{\partial u}{\partial x}\frac{\partial \xi}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial \xi}{\partial y}+\xi f\Big]dx\,dy.\end{aligned}$$

Остается проверить, что третье слагаемое в правой части обращается в нуль. Действительно, из очевидных тождеств

$$\frac{\partial u}{\partial x}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial \xi}{\partial y} \equiv \frac{\partial}{\partial x}\Big(\xi\frac{\partial u}{\partial x}\Big) + \frac{\partial}{\partial y}\Big(\xi\frac{\partial u}{\partial y}\Big) - \xi\Big(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\Big) \equiv$$
$$\equiv \frac{\partial}{\partial x}\Big(\xi\frac{\partial u}{\partial x}\Big) + \frac{\partial}{\partial y}\Big(\xi\frac{\partial u}{\partial y}\Big) - \xi f$$

следует, что

$$\iint\limits_D \Big(\frac{\partial u}{\partial x}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial \xi}{\partial y} + \xi f\Big)dx\,dy =$$
$$= \iint\limits_D \Big[\frac{\partial}{\partial x}\Big(\xi\frac{\partial u}{\partial x}\Big) + \frac{\partial}{\partial y}\Big(\xi\frac{\partial u}{\partial y}\Big)\Big]dx\,dy = \int\limits_\Gamma \xi\frac{\partial u}{\partial n}\,ds = 0, \qquad (5)$$

где $\partial/\partial n$ — производная по внутренней нормали.

В предпоследнем переходе в цепочке равенств (5) мы воспользовались теоремой из векторного анализа, в силу которой интеграл от дивергенции векторного поля по области равен потоку этого векторного поля через границу области. В данном случае этот поток $\int\limits_\Gamma \xi\frac{\partial u}{\partial n}\,ds$ обращается в нуль, так как $\xi|_\Gamma = 0$. □

Таким образом, задача (A) допускает следующую вариационную постановку: среди всех функций класса W, удовлетворяющих условию (2), найти ту, которая придает наименьшее значение функционалу $I(w)$, определенному формулой (3).

2. Сходимость минимизирующих последовательностей. Точным решением задачи (A), как мы видели, является та функция $w(x,y) \equiv u(x,y)$, которая придает наименьшее значение функционалу $I(w)$ среди всех допустимых функций, т. е. функций $w \in W$, удовлетворяющих условию $w|_\Gamma = \varphi(s)$.

Численное решение вариационной задачи об отыскании $u(x,y)$ состоит в построении такой функции $w(x,y)$, которая придает функционалу $I(w)$ хотя и не наименьшее значение $I(u)$, но "близкое" к наименьшему. Точнее, для вычисления должен быть указан способ построения членов такой последовательности допустимых функций $w_N(x,y) \in W(w_N|_\Gamma = \varphi(s))$, для которой

$$\lim_{N\to\infty} I(w_N) = I(u).$$

Такая последовательность допустимых функций называется *минимизирующей.* Выбирая член $w_N(x,y)$ минимизирующей последовательности с достаточно большим номером N, можно добиться того, чтобы $I(w_N)$ отличалось от $I(u)$ сколь угодно мало.

3. Вариационный метод Ритца. Естественно ожидать и можно строго показать, что приближением для решения u задачи Дирихле могут служить те функции w из числа допустимых ($w \in W$, $w|_\Gamma =$

$= \varphi(s))$, на которых функционал $I(w)$ принимает значения, близкие к минимальному значению $I(u)$.

Для фактического отыскания приближенных решений Ритцем в 1908 г. был предложен прием, который мы изложим применительно к задаче (А). Для удобства изложения будем считать, что $\varphi(s) = 0$, так что $u|_\Gamma = 0$.

Формальная схема отыскания приближенного решения по методу Ритца состоит в следующем. Обозначим через $\overset{\circ}{W}$ линейное подпространство всех функций $w \in W$, удовлетворяющих граничному условию $w|_\Gamma = 0$. Задаем натуральное N и фиксируем какие-нибудь N линейно независимых функций

$$\omega_1^{(N)}(x,y),\ \omega_2^{(N)}(x,y),\ \ldots,\ \omega_N^{(N)}(x,y), \tag{6}$$

удовлетворяющих условию

$$\omega_n^{(N)}\big|_\Gamma = 0, \quad n = 1, 2, \ldots, N. \tag{7}$$

Рассмотрим теперь N-мерное линейное пространство $\overset{\circ}{W}{}^{(N)}$ всевозможных линейных комбинаций функций (6):

$$w_N(x, y, a_1, a_2, \ldots, a_N) = \sum_n a_n w_n^{(N)},$$

где $a_1, a_2, \ldots, a_N$ — произвольные вещественные числа. Будем искать теперь вместо функции $w \equiv u$, придающей минимум функционалу $I(w)$ на пространстве $\overset{\circ}{W}$, такую функцию $w_N(x, y, a_1, a_2, \ldots, a_N)$, которая придает минимум функционалу $I(w)$ на множестве всех функций из N-мерного пространства $\overset{\circ}{W}{}^{(N)}$. Эту функцию $w(x, y, a_1, a_2, \ldots$ $\ldots, a_N) \equiv w_N(x, y)$ и примем за приближенное решение при сделанном выборе N базисных функций (6). Задача об отыскании функций $w_N(x, y)$ несравненно проще задачи отыскания точного решения $u(x, y)$.

Действительно,

$$I(w_N) = \iint\limits_D \left[\left(\frac{\partial}{\partial x}\sum_n a_n\omega_n^{(N)}\right)^2 + \left(\frac{\partial}{\partial y}\sum_n a_n\omega_n^{(N)}\right)^2\right]dx\,dy + \\ + 2\iint\limits_D f\left(\sum_n a_n\omega_n^{(N)}\right)^2 dx\,dy, \tag{8}$$

и речь идет об отыскании N чисел $a_1, a_2, \ldots, a_N$, придающих минимум функции $I[w_N(x, y, a_1, a_2, \ldots, a_N)]$ от N переменных.

Первое слагаемое в правой части (8) является квадратичной формой от $a_1, a_2, \ldots, a_N$. Ввиду линейной независимости системы функций (6) эта форма положительно определена, а в силу этого квадратичная функция (8) имеет единственный минимум.

Этот минимум достигается при тех значениях $a_n = \widetilde{a}_n$ $(n = 1, 2, \ldots$ $\ldots, N)$, при которых

$$\frac{\partial I[w_N(x, y, a_1, a_2, \ldots, a_N)]}{\partial a_n} = 0, \quad n = 1, 2, \ldots, N. \tag{9}$$

Подробно эта линейная система может быть записана в виде

$$\sum_{i=1}^{N} a_i \iint\limits_{D} \Big[\frac{\partial \omega_i^{(N)}}{\partial x} \frac{\partial \omega_n^{(N)}}{\partial x} + \frac{\partial \omega_i^{(N)}}{\partial y} \frac{\partial \omega_n^{(N)}}{\partial y}\Big] dx\, dy = \\ = -\iint\limits_{D} f \omega_n^{(N)} dx\, dy, \quad n = 1, 2, \ldots, N. \tag{10}$$

Для более краткой записи полученной системы в дальнейшем наряду с нормированным пространством W рассмотрим линейное пространство $\widetilde{W}$, состоящее из тех же функций, что и W, но со скалярным умножением (w', w''):

$$(w', w'') = \iint\limits_{D} \Big(\frac{\partial w'}{\partial x} \frac{\partial w''}{\partial x} + \frac{\partial w'}{\partial y} \frac{\partial w''}{\partial y}\Big) dx\, dy + \int\limits_{\Gamma} \sigma(s) w' w'' ds, \tag{11}$$

где $\sigma(s) \geqslant \sigma_0 = \text{const} > 0$ — какая-нибудь фиксированная функция. Это скалярное умножение индуцирует норму $\|w\|_{\widetilde{W}}$ в пространстве $\widetilde{W}$ по формуле

$$\|w\|^2_{\widetilde{W}} = (w, w). \tag{12}$$

Обозначим через $\overset{\circ}{\widetilde{W}}$ подпространство функций $w \in \widetilde{W}$, удовлетворяющих условию $w|_\Gamma = 0$.

После введения скалярного умножения система (10) благодаря условию $w_n^{(N)}(x, y)|_\Gamma = 0$ примет вид

$$\sum_{i=1}^{n} a_i (\omega_i^{(N)}, \omega_n^{(N)}) = -\iint\limits_{D} f \omega_n^{(N)} dx\, dy, \quad n = 1, 2, \ldots, n. \tag{13}$$

Заметим, что матрица системы (13)

$$\omega^N = \begin{bmatrix} (\omega_1^{(N)}, \omega_1^{(N)}) & \ldots & (\omega_1^{(N)}, \omega_N^{(N)}) \\ \ldots & \ldots & \ldots \\ (\omega_N^{(N)}, \omega_1^{(N)}) & \ldots & (\omega_N^{(N)}, \omega_N^{(N)}) \end{bmatrix} \tag{14}$$

есть матрица Грама системы линейно независимых функций (6). Как известно из курса линейной алгебры, ее определитель отличен от нуля.

Решение $a_n = \widetilde{a}_n$ $(n = 1, 2, \ldots, N)$ системы (13) и позволяет опреде-

лить функцию

$$\widetilde{w}_N(x,y) = w_N(x,y,\widetilde{a}_1,\widetilde{a}_2,...,\widetilde{a}_N),$$

которую принимают за приближенное решение.

4. Пример вариационно-разностной схемы. Фиксируем натуральное N. Впишем в контур Γ, ограничивающий область D, замкнутую несамопересекающуюся ломаную $Q_1^N Q_2^N ... Q_m^N Q_1^N$ с вершинами в некоторых точках Q_j^N. Обозначим многоугольник $Q_1^N Q_2^N ... Q_m^N Q_1^N$ через D_N.

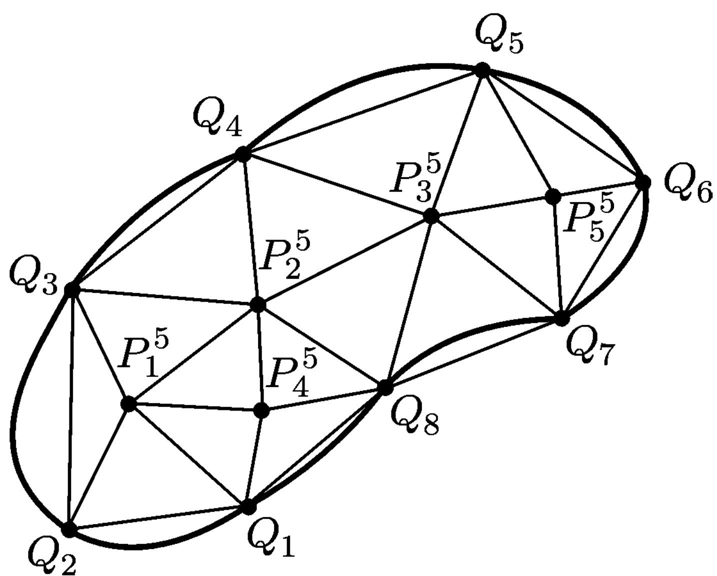


Рис. 40

Разобьем многоугольник D_N на треугольники так: чтобы каждое звено ломаной $Q_1^N Q_2^N ... Q_m^N Q_1^N$ оказалось стороной одного из треугольников; чтобы каждые два треугольника этого разбиения либо не пересекались, либо имели общую вершину, либо имели общую сторону; чтобы общее число вершин P_1^N, $P_2^N, ..., P_N^N$ этих треугольников, лежащих внутри многоугольника D_N, равнялось N. Совокупность точек P_j^N и примем за сетку (рис. 40).

Теперь построим базисные функции $\omega_1^N, \omega_2^N, ..., \omega_N^N$. Определим базисную функцию $\omega_n^N(x,y)$ $(n = 1,2,...,N)$. Сначала зададим ее в точках сетки по формуле

$$\omega_n^N(P_k^N) = \delta_n^k, \quad n,k = 1,2,...,N. \tag{15}$$

Затем зададим ее в точках $Q_1^N, Q_2^N, ..., Q_N^N$, положив равной нулю в этих точках. Таким образом, функция $\omega_n^N(x,y)$ уже определена во всех вершинах треугольников, образующих разбиение D_N. В каждом из этих треугольников доопределим ее линейно. Осталось определить ее в области $D \backslash D_N$, где положим ее равной нулю.

Заметим, что в тех треугольниках, для которых P_n^N не является одной из вершин, функция $\omega_n^N(x,y)$ при нашем построении окажется равной нулю. В треугольнике с вершиной в точке P_n^N функция $\omega_n^N(x,y)$ в пространстве (x,y,w) изобразится куском плоскости, проходящей через сторону, лежащую против вершины P_n^N и приподнятую на единицу над точкой P_n^N. Система уравнений Ритца (13) для определения коэффициентов $a_n = \widetilde{a}_n$, задающих приближенное решение

$$w_N = \sum_{i=1}^{N} w_N(P_i^N)\omega_i^N(x,y),$$

имеет вид

$$\sum_{i=1}^{N} \omega_N(P_i^N)(\omega_i^N, \omega_n^N) = -\iint\limits_D f\omega_n^N\,dx\,dy, \quad n = 1,2,...,N. \tag{16}$$

Это и есть вариационно-разностная схема, или схема метода конечных элементов, возникающая при сделанном выборе сетки и базисных функций.

Матрица (14) разностной схемы (16) имеет элементами числа

$$(\omega_n^N, \omega_i^N) = \iint\limits_D \Big[\frac{\partial \omega_n^N}{\partial x}\,\frac{\partial \omega_i^N}{\partial x} + \frac{\partial \omega_n^N}{\partial y}\,\frac{\partial \omega_i^N}{\partial y}\Big]dx\,dy. \tag{17}$$

Очевидно, что отличаться от нуля могут только те числа (ω_n^N, ω_i^N), для которых точки P_n^N, P_i^N являются вершинами одного и того же треугольника разбиения. Действительно, если P_n^N, P_i^N не являются соседними точками сетки в этом смысле, то области, в которых $\omega_n^N \not\equiv 0$, $\omega_i^N \not\equiv 0$, не пересекаются, а потому подынтегральное выражение в формуле (17) всюду в области интегрирования тождественно обращается в нуль.

Таким образом, n-е уравнение из числа образующих вариационно-разностную схему (16) связывает значения искомой функции в точке P_i^N со значениями этой функции только в тех точках сетки, которые являются для нее соседними.

Вычисление коэффициентов по формуле (17) трудностей не вызывает. Действительно, коэффициент (ω_n^N, ω_i^N) представляет собой интеграл от величины

$$\frac{\partial \omega_n^N}{\partial x}\,\frac{\partial \omega_i^N}{\partial x} + \frac{\partial \omega_n^N}{\partial y}\,\frac{\partial \omega_i^N}{\partial y}, \tag{18}$$

распространенный по паре треугольников разбиения, имеющих отрезок $P_n^N P_i^N$ своей общей стороной.

Далее, интеграл по любому одному из этих треугольников полностью определяется длинами его сторон, но не зависит от поворотов и сдвигов этого треугольника. В самом деле, постоянная внутри треугольника величина (8) подынтегрального выражения представляет собой произведение длин векторов $\operatorname{grad}\omega_n^N$ и $\operatorname{grad}\omega_i^N$ на косинус угла между этими векторами и выражается формулой

$$\frac{1}{h_n h_i}\cos(\overline{h}_n, \overline{h}_i). \tag{19}$$

Здесь h_n, h_i — длины высот, выходящих из вершин P_n^N, P_i^N соответственно, а $\overline{h}_n$, $\overline{h}_i$ — векторы, направленные вдоль соответствующих высот в сторону вершин. Интеграл по треугольнику получается умножением величины (19) на площадь треугольника.

Построение вариационно-разностной схемы (16) при сделанном выборе точек Q_n^N, P_n^N закончено.

§ 3. Вычисление решений сеточных аналогов краевых задач

Изложим и сопоставим некоторые употребительные способы вычисления решений больших систем линейных алгебраических уравнений, возникающих при дискретизации эллиптических краевых задач с помощью простейших разностных схем на регулярных сетках или схем метода конечных элементов, допускающих существенно нерегулярные сетки.

1. Точное решение разностного аналога задачи Дирихле для уравнения Пуассона в квадратной области. Разностная схема, о которой идет речь, изложена в § 1. В случае однородных краевых условий, т. е. $u|_\Gamma = 0$, решение можно выписать в виде конечного ряда Фурье. Соответствующая формула приведена в § 7 из гл. 4. В случае если число N $(h = 1/N)$ точек сетки является степенью двойки: $N = 2^k$, $k = \lg_2 N$, вычисления можно организовать так, чтобы потребовалось всего $O(N^2 \ln N)$ арифметических операций. Обычно в математическом обеспечении ЭВМ имеется программа, реализующая соответствующий алгоритм, который называют *быстрым преобразованием Фурье* (БПФ).

В случае $\psi(s) \neq 0$ также можно воспользоваться методом Фурье. Для этого достаточно заметить, что значения решения u_{mn} во внутренних точках $(mh, nh) \in D_h^0$ совпадают с решением разностной задачи Дирихле с нулевыми условиями на границе и с правой частью $\widetilde{\varphi}_{mn}((mh, nh) \in D_h^0)$, которая совпадает с φ_{mn} в тех точках, которые более чем на h отстоят от границы, а в приграничных точках $\widetilde{\varphi}_{mn}$ определяется формулой

$$\widetilde{\varphi}_{mn} = \varphi_{mn} - \frac{1}{h^2}[\widetilde{\psi}(s_{m+1,n}) + \widetilde{\psi}(s_{m-1,n}) + \widetilde{\psi}(s_{m,n+1}) + \widetilde{\psi}(s_{m,n-1})],$$

$$\widetilde{\psi}(s_{m\pm k,n\pm l}) = \begin{cases} 0, & (m \pm k, n \pm l) \in D_h^0, \\ \psi(s_{m\pm k,n\pm l}), & (m \pm k, n \pm l) \in \Gamma_h. \end{cases}$$

2. Методы простой итерации, трехслойные методы Чебышева и методы сопряженных градиентов. В гл. 5 мы рассмотрели названные методы для решения абстрактных систем линейных уравнений вида

$$Ax = f; \quad A\colon R^n \to R^n, \quad A = A^* > 0.$$

Пусть $\mu(A)$ — число обусловленности оператора A. Были установлены формулы, показывающие, что для уменьшения погрешности ε_0 нулевого приближения x_0 в заданное число $\varepsilon > 0$ раз требуется найти приближение x_p с номером p, который выражается соответственно формулами вида

$$p_1 = p_1(\mu, \varepsilon) = O(\mu \ln \varepsilon),$$

$$p_2 = p_2(\mu, \varepsilon) = O(\sqrt{\mu} \ln \varepsilon),$$
$$p_3 = p_3(\mu, \varepsilon) = O(\sqrt{\mu} \ln \varepsilon).$$

При этом первые два метода вычислительно устойчивы, а метод сопряженных градиентов при больших значениях числа $\mu = \mu(A)$ может оказаться неустойчивым.

Для ускорения сходимости можно воспользоваться эквивалентными по спектру, или энергетически эквивалентными, операторами. Оператор $B = B^* > 0$ *энергетически эквивалентен* оператору A с постоянными $\gamma_1 > 0$, $\gamma_2 > 0$, если

$$\gamma_1(Bx, x) \leqslant (Ax, x) \leqslant \gamma_2(Bx, x).$$

Если решение задачи вида $Bx = \varphi$ легко вычисляется, а $\gamma_2/\gamma_1 \ll \mu(A)$, то имеет смысл перейти от системы $Ax = f$ к системе $Cx = \psi$, где $C = B^{-1}A$, $\psi = B^{-1}f$, и решать эту последнюю итерациями. Ускорение сходимости будет достигнуто за счет того, что $\mu_B(C) \leqslant \gamma_2/\gamma_1 \ll \mu$.

Все эти соображения и факты можно использовать при вычислении решений систем, возникающих при дискретизации эллиптических задач. Свойство $A = A^* > 0$ для них часто имеет место.

Для системы, возникающей в методе конечных элементов, это свойство вытекает из вариационного происхождения этой системы, так как матрица системы есть матрица Грама ω^N системы линейно независимых элементов из § 2.

Ускорение сходимости за счет выбора оператора B легко получить для задачи Дирихле в случае эллиптического уравнения с переменными коэффициентами в квадратной области, используя в качестве B оператор для задачи с постоянными коэффициентами, для обращения которого можно воспользоваться алгоритмом быстрого преобразования Фурье.

В случае более сложных областей, краевых условий или нерегулярных сеток ускорение за счет построения оператора B, который был бы легко обратим и для которого $\mu_B(C) \ll \mu(A)$, $C = B^{-1}A$, затруднительно.

Подробное изложение итерационных методов сеточных аналогов краевых задач см. в [7, 20].

В настоящее время широкое распространение и развитие (особенно в связи с уравнениями метода конечных элементов на нерегулярных сетках) получил многосеточный метод Федоренко, которому мы посвятим следующий параграф.

§ 4. Многосеточный метод Федоренко

В работе Р.П. Федоренко (см.: ЖВМ и МФ.—1961.—Т. I, № 5) предложен метод итерационного решения разностных эллиптических задач, названный *релаксационным.* Для уменьшения нормы первона-

чальной погрешности вдвое этот метод требует всего CN арифметических действий, где $C = \text{const}$, а N — число точек в области. Очевидно, что по порядку числа арифметических действий этот метод неулучшаем, так как число расчетных точек также $O(N)$.

Границы применимости метода Федоренко почти такие же, как у метода простых итераций. Регулярность сетки не предполагается. Дополнительным ограничением является требование "плавности", "гладкости" первых собственных функций оператора задачи, которое для эллиптических задач обычно выполнено.

Цитированная работа Федоренко была развита в ряде последующих работ многих авторов.

Обоснования многосеточного метода Федоренко довольно громоздки, поэтому мы ограничимся качественным описанием идеи метода и самого алгоритма Федоренко в простом случае, отсылая за доказательствами и общими конструкциями к оригинальным работам и [24].

1. Идея метода. При решении итерациями задачи

$$\begin{gathered} \Delta_h u_{mn} - \varphi_{mn} = 0, \quad m, n = 1, 2, ..., M-1, \\ u_{mn}\big|_{\Gamma_h} = \psi(s_{mn}) \end{gathered} \tag{1}$$

будем отправляться от метода простых итераций

$$\begin{gathered} u_{mn}^{p+1} = u_{mn}^p + \tau(\Delta_h u_{mn}^p - \varphi_{mn}), \quad m, n = 1, 2, ..., M-1, \\ p = 0, 1, ..., \\ u_{mn}^{p+1}\big|_{\Gamma_h} = \psi(s_{mn}), \quad [u_{mn}^0] \text{ задано}, \end{gathered} \tag{2}$$

который в целом сходится очень медленно, но неравномерно на различных гармониках. Погрешность $\varepsilon^p = u^p - u$ в соответствии с § 7 гл. 4 записывается в виде конечного ряда Фурье

$$\varepsilon^p = \sum [\lambda_{rs}(\tau)]^p c_{rs}^{(0)} \psi^{(r,s)}, \tag{3}$$

где $c_{rs}^{(0)}$ — коэффициенты разложения погрешности $\varepsilon^0 = u - u^0$ нулевого приближения, $\lambda_{rs} = 1 - 4\tau M^2[\sin^2(\pi r/(2M)) + \sin^2(\pi s/(2M))]$. Числа λ_{rs} лежат на отрезке $\lambda_{\text{лев.}} \leqslant \lambda \leqslant \lambda_{\text{прав.}}$, где

$$\lambda_{\text{лев.}} = \lambda_{M-1,M-1} \approx 1 - 8\tau M^2, \quad \lambda_{\text{прав.}} = \lambda_{11} \approx 1 - 2\pi^2\tau.$$

Положим

$$\tau = 3/(16M^2). \tag{4}$$

Если при этом условии хотя бы одно из чисел r или s больше, чем $M/2$, то

$$|\lambda_{rs}| < 0{,}6. \tag{5}$$

Поэтому вклад высокочастотных гармоник $\psi^{(r,s)}$ ($r \geqslant M/2$ или $s \geqslant M/2$) в погрешность (3) за один шаг итерационного процесса уменьшается почти вдвое и вскоре становится малым. После нескольких

итераций по формуле (2) погрешность будет в основном состоять только из гладких компонент (гармоники $\psi^{(r,s)}$, $r < M/2$, $s < M/2$), потому что низкочастотные гармоники $\psi^{(r,s)}$ умножаются на числа λ^p_{rs}, которые ближе к единице. Очень медленно гасится вклад первой гармоники $\psi^{(1,1)}$ при сделанном выборе τ, так как

$$\lambda_{11} \approx 1 - \frac{3\pi^2}{8M^2} \; (\approx 1). \tag{6}$$

Обозначим полученное в процессе итераций (2) приближение u^p через U, а погрешность $\varepsilon^p = u^p - u = U - u$ через v. Если бы мы знали погрешность v, то нашли бы искомое решение $u = U - v$. Однако относительно v мы знаем только, что оно удовлетворяет уравнению

$$\Delta_h v = \xi, \quad v\big|_{\Gamma_h} = 0, \tag{7}$$

где известная сеточная функция ξ — это невязка, возникающая при подстановке $u^p = U$ в уравнение (1):

$$\Delta_h u^p - \Delta_h u = \Delta_h U - \varphi.$$

Задача (7) для определения поправки v проще исходной задачи (1) лишь в том отношении, что заранее известно, что v — гладкая сеточная функция. Поэтому для определения v вместо задачи (7) можно приближенно рассматривать такую же задачу на вдвое более крупной сетке, которая (при четном M) является подсеткой исходной сетки:

$$\Delta_{2h} v^* = \xi^*, \quad v^*\big|_{\Gamma_{2h}} = 0. \tag{1*}$$

Звездочкой мы обозначили величины на укрупненной сетке. Задачу (1*) решаем итерациями по формулам

$$\begin{gathered} (v^*_{mn})^{p+1} = (v^*_{mn})^p + \tau^*[\Delta_{2h}(v^*_{mn})^p - \xi^*_{mn}], \\ m, n = 1, 2, ..., M^* - 1, \\ (v^*_{mn})^{p+1}\big|_{\Gamma^*_{2h}} = 0, \end{gathered} \tag{2*}$$

приняв за нулевое приближение $(v^*_{mn})^0 \equiv 0$. Здесь

$$M^* = M/2, \quad \tau^* = 4\tau.$$

Каждый шаг итераций (2*) вчетверо менее трудоемок, чем шаг итераций (2), потому что расчетных точек вчетверо меньше. Кроме того, благодаря $\tau^* = 4\tau$ быстрее происходит погашение самой медленно убывающей компоненты погрешности. В соответствии с (6)

$$\lambda^*_{11} = 1 - \frac{3\pi^2}{8(M^*)^2} = 1 - 4\,\frac{3\pi^2}{8M^2} < \lambda_{11},$$

и для погашения вклада $(\psi^*)^{(1,1)}$ в e раз нужно вчетверо меньше итераций. Результат итераций по формуле (2*) обозначим V^*. Проинтерполируем V^* на исходную сетку (линейно). Гладкие компоненты будут получены почти правильно. Возникающая при интерполяции

погрешность будет относительно интерполируемой гладкой функции мала, но ее разложение Фурье будет содержать все гармоники (погрешность интерполяции негладкая из-за изломов в узлах интерполяции). Кроме того, негладкая компонента, не имеющая отношения к искомой поправке, при интерполяции тоже дает случайный вклад в негладкую составляющую полученной при интерполяции функции V. Итак, гладкая компонента разности $U - V$ близка к гладкой компоненте искомого решения $u = U - v$, но негладкая компонента не очень мала и носит случайный характер.

Поэтому следует проделать еще несколько шагов исходного итерационного процесса (2), приняв $U - V$ за начальное приближение. Это погасит привнесенную при интерполяции негладкую составляющую погрешности, которую итерации (2) уменьшают почти вдвое за один шаг.

2. Описание алгоритма. При большом M (мелкой сетке) задача (1*) на укрупненной сетке все еще трудоемка. Поэтому при ее решении в свою очередь целесообразно проделать еще одно укрупнение сетки вдвое, а при решении задачи на вчетверо укрупненной сетке снова использовать процесс удвоения сетки и увеличения h и т. д. Будем считать для простоты, что $M = 2^k$, т. е. M является некоторой степенью двойки.

На исходной сетке делаем несколько шагов итераций (2), чтобы “сгладить” погрешность. Погрешность нам неизвестна, поэтому можно следить за этим по невязке $\Delta_h u^p - \varphi$, которая тоже сглаживается. Результат вычислений $U = u^p$ запоминаем. Затем для поправки v рассматриваем задачу на укрупненной сетке, делаем несколько итераций (2*), чтобы сгладить “поправку к поправке”, и результат V^* запоминаем (он занимает вчетверо меньше места в памяти, чем U). Для вычисления поправки к V^* рассматриваем задачу на еще вдвое укрупненной сетке, делаем несколько итераций с шагом $\tau^{**} = 4\tau^* = 16\tau$ и запоминаем результат V^{**}. Этот процесс вычисления поправок к поправкам на вдвое укрупненных сетках продолжается k раз, пока не дойдет до самой крупной сетки и поправки $\widetilde{V}^{(k*)}$.

Затем начинаем возвращение на мелкую сетку. Сначала с самой крупной сетки интерполируем полученную там последнюю поправку $\widetilde{V}^{(k*)}$ на вдвое более мелкую сетку, вносим эту проинтерполированную поправку в $V^{(k-1)*}$ и делаем несколько итераций, чтобы погасить привнесенную при интерполяции погрешность. Результат этих итераций интерполируем на еще вдвое более мелкую сетку, уточняем с его помощью хранящуюся для этой сетки поправку $V^{(k-2)*}$, делаем несколько итераций и производим следующую интерполяцию. На предпоследнем шаге после внесения в $\widetilde{V}^*$ поправки и итераций получим поправку V^*, которую интерполируем на исходную сетку. Проделав несколько итераций (2) над $u - V$, получим результат.

ЧАСТЬ IV

МЕТОДЫ ГРАНИЧНЫХ УРАВНЕНИЙ ДЛЯ ЧИСЛЕННОГО РЕШЕНИЯ КРАЕВЫХ ЗАДАЧ

Разностные методы численного решения краевых задач обладают весьма широкой областью применимости. Тем не менее существенные трудности для использования разностных методов возникают при разностной аппроксимации краевых условий общего вида (в частности, нелокальных) в областях произвольной формы, поскольку в результате этой аппроксимации должна получаться устойчивая и удобная для численного решения разностная схема. Существенную трудность для применения разностных методов представляют также внешние задачи с условиями на бесконечности.

Эти трудности в ряде случаев удается преодолеть, используя редукцию исходной краевой задачи относительно неизвестной функции в области к некоторым уравнениям относительно других неизвестных функций, определенных на границе этой области. К тому же при такой редукции на единицу уменьшается геометрическая размерность задачи (граница двумерной области — линия). Мы познакомимся с двумя способами такой редукции, а также и с соответствующими методами численного решения уравнений на границе.

Первый метод редукции исходной задачи к уравнениям на границе есть метод граничных интегральных уравнений (ГИУ) классической теории потенциала, идущей от работ Фредгольма начала века. Дискретизация и численное решение возникающих граничных интегральных уравнений осуществляются с помощью специальных квадратурных формул, или так называемого метода граничных элементов (МГЭ).

Второй метод редукции исходной задачи приводит к таким уравнениям на границе, в структуру которых входит оператор проектирования (он был предложен Кальдероном в 1963 г.). Мы излагаем некоторую модификацию граничных уравнений с проекторами (ГУРП) Кальдерона, предпринятую для того, чтобы ГУРП стали удобны для дискретизации и численного решения методом разностных потенциалов (МРП), предложенным в 1969 г.

ГУРП и МРП можно применять во многих случаях, когда ГИУ и МГЭ неприменимы. Мы ограничимся изложением идей, лежащих в основе названных методов, и описанием границ применимости этих

методов, а также укажем литературу, где можно найти более полное изложение.

ГЛАВА 12

ГРАНИЧНЫЕ ИНТЕГРАЛЬНЫЕ УРАВНЕНИЯ И МЕТОД ГРАНИЧНЫХ ЭЛЕМЕНТОВ ДЛЯ ИХ ЧИСЛЕННОГО РЕШЕНИЯ

Познакомимся со способами перехода от краевых задач к интегральным уравнениям на границе, с приемом, лежащим в основе дискретизации интегральных уравнений, а также укажем трудности, ограничивающие класс задач, для которого применение метода достаточно естественно.

§ 1. Способы редукции краевых задач к ГИУ

Для выяснения основных идей ограничимся рассмотрением внутренних и внешних задач Дирихле и Неймана для уравнения Лапласа

$$\Delta u \equiv \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2} = 0.$$

Пусть D — ограниченная область в трехмерном пространстве с достаточно гладкой границей Γ. Рассмотрим следующие четыре задачи:

$$\Delta u = 0, \quad x \in D, \quad u|_\Gamma = \varphi(x)|_\Gamma; \tag{1}$$

$$\Delta u = 0, \quad x \in D, \quad \partial u/\partial n|_\Gamma = \varphi(x)|_\Gamma; \tag{2}$$

$$\Delta u = 0, \quad x \notin D, \quad u|_\Gamma = \varphi(x)|_\Gamma; \tag{3}$$

$$\Delta u = 0, \quad x \notin D, \quad \partial u/\partial n|_\Gamma = \varphi(x)|_\Gamma. \tag{4}$$

Здесь $\varphi(x)|_\Gamma$ — заданная функция.

Задачи (1), (3) суть внутренняя и внешняя задачи Дирихле, а задачи (2) и (4) — внутренняя и внешняя задачи Неймана.

Решение внешних задач будем искать в классе функций $u(x)$, стремящихся к нулю при $r \to \infty$, где $r = (x_1^2 + x_2^2 + x_3^2)^{1/2}$.

В математической физике доказывается, что внутренняя и внешняя задачи Дирихле, а также внешняя задача Неймана всегда имеют одно и только одно решение. Внутренняя задача Неймана имеет решение только в том случае, если интеграл $\int \varphi \, ds$ по поверхности Γ

обращается в нуль:

$$\int_\Gamma \varphi\, ds = 0, \tag{5}$$

где ds — элемент площади на поверхности Г. В этом случае решение определено с точностью до произвольного постоянного слагаемого.

Для перехода от задач (1)–(4) к интегральным уравнениям используется функция

$$G(x) = -\frac{1}{4\pi}\,\frac{1}{r}. \tag{6}$$

Эта функция является фундаментальным решением уравнения Лапласа в трехмерном пространстве. Это означает, что стремящееся к нулю при $r \to \infty$ решение уравнения

$$\Delta u = f(x)$$

с произвольной достаточно гладкой правой частью $f(x)$, обращающейся в нуль вне какой-либо сферы, можно записать формулой

$$u(x) = \iiint G(x-y) f(y)\, dy_1\, dy_2\, dy_3, \tag{7}$$

где интегрирование распространено по всему трехмерному пространству.

Для перехода от задач (1)–(4) к интегральным уравнениям на границе используются потенциал простого слоя

$$V(x) = \int_\Gamma G(x-y)\rho(y)\, ds_y \tag{8}$$

и потенциал двойного слоя

$$W(x) = \int_\Gamma \sigma(y)\, \frac{\partial G(x-y)}{\partial n_y}\, ds_y. \tag{9}$$

Здесь ds_y — элемент площади на поверхности Г, ограничивающей область D, n_y — направление внешней по отношению к D нормали в точке $y \in \Gamma$. Функции $\rho(y)$ и $\sigma(y)$ называются соответственно *плотностями потенциалов простого и двойного слоев.*

Известно (и легко видеть), что $G(x)$ (а значит, и потенциалы (8), (9) при $x \notin \Gamma$) удовлетворяет уравнению Лапласа при $x \neq 0$. Таким образом, потенциалы (8), (9) суть семейства гармонических функций, зависящих от произвольных плотностей $\rho(y)$, $\sigma(y)$, $y \in \Gamma$.

Будем искать решения задач (1), (3) в виде потенциалов двойного слоя (9), а решения задач (2), (4) в виде потенциалов простого слоя (8). В курсах уравнений математической физики показано, что возникают следующие интегральные уравнения:

1) для внутренней задачи Дирихле (1)

$$\sigma(x) + \frac{1}{2\pi}\int\limits_{\Gamma}\sigma(y)\,\frac{\partial G}{\partial n_y}\,dS_y = -\frac{1}{2\pi}\,\varphi(x), \quad x \in \Gamma; \tag{10}$$

2) для внешней задачи Неймана (4)

$$\rho(x) - \frac{1}{2\pi}\int\limits_{\Gamma}\rho(y)\,\frac{\partial G}{\partial n_y}\,dS_y = \frac{1}{2\pi}\,\varphi(x), \quad x \in \Gamma; \tag{11}$$

3) для внутренней задачи Неймана (2)

$$\rho(x) + \frac{1}{2\pi}\int\limits_{\Gamma}\rho(y)\,\frac{\partial G}{\partial n_y}\,dS_y = -\frac{1}{2\pi}\,\varphi(x), \quad x \in \Gamma; \tag{12}$$

4) для внешней задачи Дирихле (3)

$$\sigma(x) - \frac{1}{2\pi}\int\limits_{\Gamma}\sigma(y)\,\frac{\partial G}{\partial n_y}\,dS_y = \frac{1}{2\pi}\,\varphi(x), \quad x \in \Gamma. \tag{13}$$

В теории потенциала устанавливается, что уравнения (10), (11) имеют и притом единственные решения $\sigma(y)$, $\rho(y)$ при произвольной функции $\varphi(x)$ $(x \in \Gamma)$. Уравнение (12) имеет решение лишь в случае выполнения условия (5). Это связано с существом исходной задачи (2), которая разрешима лишь при этом условии. Однако уравнение (13) также разрешимо не при произвольной функции $\varphi(x)$, хотя исходная внешняя задача Дирихле (3) разрешима всегда. Переход к интегральному уравнению (13) в этом случае оказался непригоден.

В теории потенциала показывается, что потеря разрешимости уравнения (13) для внешней задачи Дирихле связана с тем, что внутренняя задача Неймана для уравнения Лапласа в случае $\varphi(x) \equiv 0$ имеет нетривиальное (ненулевое) решение. Это обстоятельство называют ***резонансом внутренней области.*** В данном случае вместо (13) легко построить другое интегральное уравнение, которое равносильно исходной задаче, но мы здесь этого делать не будем.

Пусть вместо уравнения Лапласа рассматривается уравнение Гельмгольца

$$\Delta u + \mu^2 u = 0, \quad \mu > 0,$$

с условием излучения на бесконечности. Соответственно

$$G(x) = -\frac{1}{4\pi}\,\frac{e^{i\mu r}}{r}.$$

Тогда вид потенциалов простого и двойного слоев, а также интегральных уравнений (10)–(13) будет тот же (только $G(x)$ другое). Внешняя задача Дирихле и внешняя задача Неймана каждая будет иметь единственное решение при любой $\varphi(x)$ $(x \in \Gamma)$ независимо от μ.

Однако интегральное уравнение вида (13) для внешней задачи Дирихле будет иметь решение при произвольной $\varphi(x)$ лишь в том случае, если при данном μ внутренняя задача Неймана при $\varphi(x) \equiv 0$ имеет только тривиальное решение, что имеет место не всегда. Аналогично обстоит дело и с разрешимостью интегрального уравнения вида (11) для внешней задачи Неймана. Оно разрешимо при любой $\varphi(x)$ лишь в том случае, если внутренняя задача Дирихле при $\varphi(x) \equiv 0$ и данном μ имеет только тривиальное решение.

Указанная трудность редукции к интегральному уравнению, связанная с наличием внутреннего резонанса для уравнения Гельмгольца, существенно затрудняет расчеты. Мы не будем останавливаться на этом подробнее.

Интегральные уравнения теории потенциала конструктивно построены в настоящее время для основных краевых задач теории упругости и некоторых других систем, для которых известны аналитические представления фундаментальных решений.

Для построения граничных интегральных уравнений наряду с теми или иными потенциалами часто используют связь между значениями решения $u|_\Gamma$ и значениями $\partial u/\partial n|_\Gamma$ на границе Γ области D. Эта связь получается из формулы Грина

$$u(x) = \int \left(G(x,y)\, \frac{\partial u}{\partial n_y} - \frac{\partial G}{\partial n_y}\, u \right) dS_y, \tag{14}$$

представляющей решение $u(x)$ во внутренней точке x через значения $u|_\Gamma$, $\partial u/\partial n|_\Gamma$.

Переходя к пределу при $x \to x_0 \in \Gamma$ и учитывая скачок потенциала двойного слоя, т. е. скачок во втором слагаемом в правой части (14), получаем интегральное соотношение, о котором идет речь. В случае задачи Неймана можно подставить значение $\partial u/\partial n|_\Gamma = \varphi(x)$ и получить разрешимое уравнение Фредгольма второго рода относительно $\partial u/\partial n|_\Gamma$. Удобство уравнения (14) по сравнению с уравнением (10) состоит в том, что вычисляется $\partial u/\partial n|_\Gamma$, а не вспомогательная плотность $\sigma(y)$.

§ 2. Граничные элементы и дискретизация ГИУ

Дискретизация ГИУ, построенных в § 1, осложнена тем, что ядра $G(x,y)$ или $\partial G/\partial n_y$ имеют особенности. Проиллюстрируем построение квадратурных формул, приближенно заменяющих интегралы, на примере уравнения (10) из § 1.

Поверхность Γ, являющаяся границей области D, разобьем на конечное число пересекающихся только, быть может, по своим границам криволинейных треугольников — граничных элементов. В каждом из граничных элементов искомую плотность будем считать постоянной. Перенумеруем граничные элементы и обозначим прибли-

женное значение плотности внутри граничного элемента Γ_j через σ_j. Заменим интеграл, входящий в уравнение (10) из § 1, суммой

$$\int\limits_{\Gamma} \sigma(y) \frac{\partial G}{\partial n_y} \, dS_y = \sum_j \sigma_j \int\limits_{\Gamma_j} \frac{\partial G}{\partial n_y} \, dS_y. \qquad (1)$$

Интегралы, входящие в правую часть (1), зависят только от Γ_j, но не от σ_j, а также от x как от параметра.

Обозначим x_k какую-либо точку граничного элемента Γ_k, а значения интегралов обозначим a_{kj}:

$$a_{kj} = \int\limits_{\Gamma_j} \frac{\partial G(x_k - y)}{\partial n_y} \, dS_y. \qquad (2)$$

Перейдем теперь от уравнения (10) из § 1 к системе линейных алгебраических уравнений

$$2\pi\sigma_k + \sum_j a_{kj}\sigma_j = -\varphi(x_k), \quad k, j = 1, 2, ..., J, \qquad (3)$$

где J — число граничных элементов.

При стремлении числа граничных элементов к бесконечности, а диаметра граничных элементов Γ_j — к нулю решение системы (3) стремится к искомой плотности $\sigma(y)$ потенциала двойного слоя, которая является решением уравнения (10) из § 1. Само искомое решение внутренней задачи Дирихле представляется с помощью потенциала двойного слоя, который можно вычислить по найденной приближенно из (3) плотности. Вычисление коэффициентов a_{kj} осуществляется точно или приближенно. В качестве граничных элементов выбираются не обязательно криволинейные треугольники. Искомая функция внутри граничного элемента не обязательно заменяется постоянной. Можно принять, что это многочлен заданного вида с неопределенными коэффициентами. Соответствующие изменения должны претерпеть уравнения (3).

Существует развитая техника построения граничных элементов, дискретизации граничных интегральных уравнений и точного или итерационного решения возникающих алгебраических систем, которая здесь не излагается.

§ 3. Область применимости ГИУ для численного решения краевых задач

В настоящее время существуют пакеты программ для численного решения многих краевых задач: для уравнений Лапласа и Гельмгольца; для системы Ламе, описывающей упругие деформации в однородном или кусочно однородном материале; для системы Стокса,

описывающей медленные течения вязкой несжимаемой жидкости, и многих других (см., например, [13, 14]).

Очевидная выгода использования ГИУ по сравнению с методом конечных разностей состоит в понижении на единицу геометрической размерности задачи за счет перехода к уравнениям на границе, а также в автоматическом учете граничных условий и условий на бесконечности. Кроме того, переход к интегральным уравнениям позволяет иногда построить алгоритмы, автоматически учитывающие гладкость входных данных и искомых функций, т. е. алгоритмы без насыщения (см. [1], а также статью: В.Н. Белых// ДАН СССР. — 1989. — Т. 304, № 3. — С. 629–631).

Главное ограничение для непосредственного использования изложенной схемы вычисления решений краевых задач с помощью ГИУ состоит в том, что для построения интегральных уравнений (а главное, для их дискретизации) требуется располагать удобным представлением ядер этих уравнений. Ядра выражаются через фундаментальные решения, которые допускают простое представление в виде формул лишь для некоторых уравнений с постоянными коэффициентами. Впрочем, метод граничных элементов во взаимодействии с различными итерационными процедурами применяют даже для некоторых нелинейных краевых задач.

Отметим еще одну трудность. Пусть фундаментальное решение известно и записано в виде простой формулы. Редукция краевой задачи к равносильному интегральному уравнению может оказаться непростым делом даже при этом условии, как мы объяснили на примере внутренних резонансов в случае задач Дирихле и Неймана для уравнений Лапласа и Гельмгольца.

ГЛАВА 13

ГРАНИЧНЫЕ УРАВНЕНИЯ С ПРОЕКТОРАМИ И МЕТОД РАЗНОСТНЫХ ПОТЕНЦИАЛОВ ДЛЯ ИХ ЧИСЛЕННОГО РЕШЕНИЯ

В гл. 12 мы рассказали об использовании классической теории потенциала для перехода от заданной краевой задачи в некоторой области к интегральным уравнениям на ее границе. Эти уравнения используют при качественном изучении и численном решении краевых задач для уравнений Лапласа, Гельмгольца, систем Ламе, Максвелла, Стокса и др.

Для редукции исходной краевой задачи к граничным интеграль-

ным уравнениям решение записывают в виде того или иного потенциала с неизвестной плотностью, сосредоточенной на границе области, относительно которой при подстановке потенциала в граничное условие и возникает интегральное уравнение. В основе построения потенциала при этом лежит фундаментальное решение исходного дифференциального уравнения или системы уравнений. От возникающих интегральных уравнений переходят к системе алгебраических уравнений, используя для замены интегралов квадратуры.

Граничные уравнения с проекторами (ГУРП), которым посвящена эта глава, как и классические граничные интегральные уравнения, осуществляют замену краевых и некоторых других задач математической физики, поставленных в области, соотношениями на границе этой области.

Прообразом общей конструкции ГУРП могут служить граничные сингулярные интегральные уравнения, использующие классическое условие Сохоцкого–Племеля для граничных значений аналитических функций (решений системы Коши–Римана).

Мы излагаем здесь ГУРП Кальдерона, конкретизированные за счет введения в конструкцию некоторых вспомогательных краевых задач. ГУРП Кальдерона основаны на задании поверхностных потенциалов без помощи фундаментального решения или функции Грина и без использования интегралов по поверхности, где сосредоточена плотность. Вместо функции Грина используется непосредственно оператор Грина, а вместо интегрального представления поверхностных потенциалов — некоторое их операторное представление.

Искомой плотностью потенциала в ГУРП Кальдерона является вектор-функция из так называемого пространства данных Коши для исследуемого эллиптического уравнения. При этом для уравнения порядка *m данными Коши на поверхности* Γ называют вектор-функцию

$$\mathbf{u}_\Gamma = \left(u|_\Gamma,\ \frac{\partial u}{\partial n}\Big|_\Gamma,\ \ldots,\ \frac{\partial^{m-1} u}{\partial n^{m-1}}\Big|_\Gamma \right),$$

заданную на поверхности Γ (n — направление нормали к Γ).

Конкретизация и модернизация конструкций Кальдерона на основе использования вспомогательной краевой задачи предпринята нами для того, чтобы получить ГУРП, приспособленные для дискретизации и численного решения широкого класса задач математической физики методом разностных потенциалов (МРП). Мы познакомим здесь читателя с ГУРП Кальдерона и МРП на примере задач для эллиптических уравнений на плоскости.

§ 1. Пример ГУРП для эллиптического уравнения второго порядка

Воспользуемся построением ГУРП для уравнения вида

$$a(x)\,\frac{\partial^2 u}{\partial x_1^2} + b(x)\,\frac{\partial^2 u}{\partial x_2^2} = 0, \quad x = (x_1, x_2) \in D, \qquad a(x) \geqslant 1, \quad b(x) \geqslant 1, \tag{1}$$

в качестве примера на котором продемонстрируем идею редукции краевых задач к равносильным ГУРП с целью их численного решения методом разностных потенциалов. Будем считать, что D — некоторая ограниченная область с достаточно гладкой границей Γ, а коэффициенты $a(x)$, $b(x)$ — достаточно гладкие функции, определенные не только в D, но и всюду на плоскости Ox_1x_2.

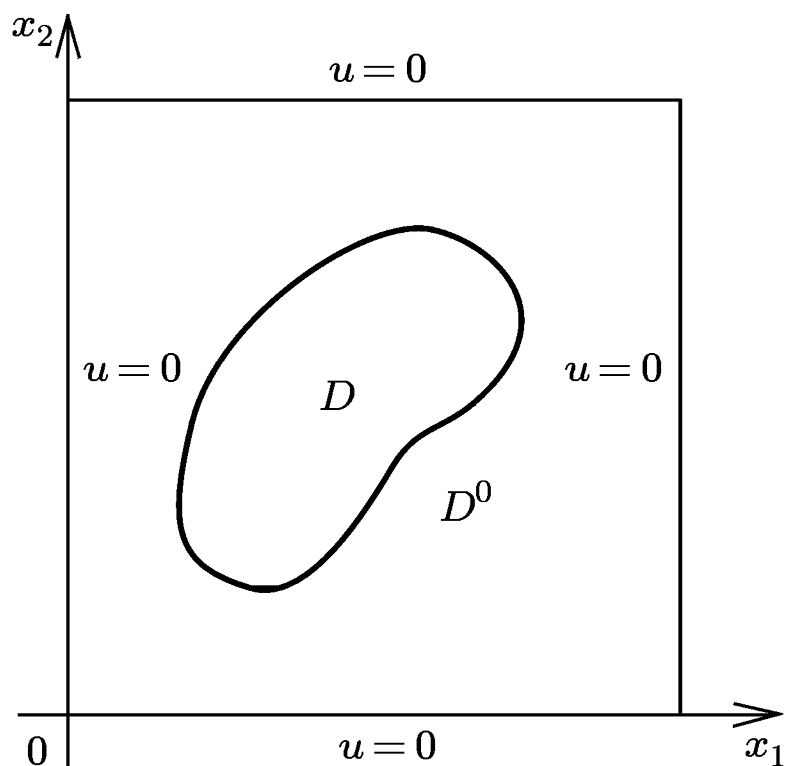


Рис. 41

1. Потенциал с плотностью из пространства данных Коши. Введем произвольно "удобную" область D^0, содержащую область D вместе с ее границей Γ. Пусть, например, областью D^0 будет служить квадрат $(0 \leqslant x_1, x_2 \leqslant 1)$. Рассмотрим следующую вспомогательную задачу Дирихле (рис. 41):

$$a(x)\,\frac{\partial^2 u}{\partial x_1^2} + b(x)\,\frac{\partial^2 u}{\partial x_2^2} = f(x), \quad x \in D^0, \tag{2}$$

$$u\big|_{\partial D^0} = 0. \tag{3}$$

Будем предполагать, что $f(x)$ принадлежит пространству F_{D^0} всех функций, определенных внутри D^0, обращающихся в нуль на Γ и удовлетворяющих условию Гёльдера

$$|f(x'') - f(x')| \leqslant c\,|x'' - x'|^{\alpha}, \quad x', x'' \in D^0, \tag{4}$$

с некоторыми постоянными $c > 0$, $\alpha \in (0, 1)$.

Введем еще линейное пространство $U_{\overline{D}^0}$ функций $u_{\overline{D}^0} = u_{\overline{D}^0}(x) = u(x)$, определенных и непрерывных на замкнутом квадрате $\overline{D}^0$, имеющих непрерывные первые производные всюду внутри D^0, включая точки границы Γ, и непрерывные вторые производные всюду внутри D^0, за исключением, быть может, точек границы Γ. Кроме того, элементы $u_{\overline{D}^0}$ пространства $U_{\overline{D}^0}$ подчиним следующему усло-

вию: функция

$$f(x) \equiv a(x)\,\frac{\partial^2 u}{\partial x_1^2} + b(x)\,\frac{\partial^2 u}{\partial x_2^2}, \quad x \notin \Gamma, \quad f(x) = 0, \quad x \in \Gamma,$$

принадлежит F_{D^0}. Из теории эллиптических уравнений известно, что при произвольной $f \in F_{D^0}$ задача (2), (3) имеет одно и только одно решение $u \in U_{\overline{D}^0}$.

Определим наконец пространство данных Коши U_Γ, относя к нему все вектор-функции вида

$$v_\Gamma = \begin{bmatrix} v_{\overline{D}}\big|_\Gamma \\ \partial v_{\overline{D}}/\partial n\big|_\Gamma \end{bmatrix}, \tag{5}$$

где $\partial/\partial n$ — производная в направлении внутренней нормали к границе Γ области D, а $v_{\overline{D}}$ — произвольная функция, которую можно доопределить всюду на $\overline{D}^0$ до некоторой функции $v_{\overline{D}^0} \in U_{\overline{D}^0}$.

Будем обозначать операцию взятия данных Коши функции $v_{\overline{D}}$ через $\mathrm{Tr}_{\Gamma\overline{D}}$ и называть ее *операцией взятия следа*, так что след $v_{\overline{D}}$ есть v_Γ:

$$v_\Gamma = \mathrm{Tr}_{\Gamma\overline{D}}\, v_{\overline{D}}. \tag{6}$$

Определение. Пусть $v_\Gamma \in U_\Gamma$ — произвольная вектор-функция из пространства данных Коши. Рассмотрим задачу вида (2), (3) с правой частью

$$f(x) = \begin{cases} 0, & \text{если} \quad x \in \overline{D}, \\ a(x)\,\dfrac{\partial^2 v}{\partial x_1^2} + b(x)\,\dfrac{\partial^2 v}{\partial x_2^2}, & \text{если} \quad x \in D^0 \backslash \overline{D}, \end{cases} \tag{7}$$

где $v(x) \equiv v_{\overline{D}^0}(x)$ есть произвольная функция из $U_{\overline{D}^0}$, данные Коши которой совпадают с v_Γ. Функцию $u_{\overline{D}}$, совпадающую с решением задачи (2), (3), (7) $u_{\overline{D}^0}$ на множестве D, будем называть *потенциалом с плотностью* v_Γ и обозначать

$$u_{\overline{D}} = P_{\overline{D}\Gamma}\, v_\Gamma. \tag{8}$$

Теорема 1. *Определение потенциала* (8) *корректно в том смысле, что решение задачи* (2), (3), (7) *существует, причем* $u_{\overline{D}}$ *зависит только от плотности* v_Γ, *но не от выбора* $v_{\overline{D}^0} \in U_{\overline{D}^0}$, *данные Коши которой совпадают с* v_Γ.

Доказательство. По определению пространства $U_{\overline{D}^0}$ для всякой функции $u_{\overline{D}^0} \in U_{\overline{D}^0}$ функция $a(x)\,\dfrac{\partial^2 u}{\partial x_1^2} + b(x)\,\dfrac{\partial^2 u}{\partial x_2^2}$ $(x \in D^0 \backslash \Gamma)$ принадлежит F_{D^0}. Но тогда правая часть (7) тоже принадлежит F_{D^0}. Таким образом, задача вида (2), (3) с правой частью (7) имеет единственное решение $u_{\overline{D}^0} \in U_{\overline{D}^0}$. Докажем, что сужение этого решения

$u_{\overline{D}^0}$ на область $\overline{D} \subset \overline{D}^0$ зависит только от v_Γ, но не от выбора $v_{\overline{D}^0}$ из числа тех, данные Коши которых совпадают с v_Γ. Ввиду линейности достаточно показать, что в случае $v_\Gamma = 0_\Gamma = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ функция $u_{\overline{D}}$, задаваемая равенством (8), есть тождественный нуль.

Действительно, в этом случае функция $u(x)$, определяемая равенством

$$u(x) = u_{\overline{D}^0}(x) = \begin{cases} 0, & \text{если} \quad x \in D, \\ v_{\overline{D}^0}(x) & \text{если} \quad x \in \overline{D}^0 \backslash \overline{D}, \end{cases} \tag{9}$$

принадлежит $U_{\overline{D}^0}$, поскольку на Γ она непрерывна вместе со своими первыми производными, которые обращаются в нуль. Остальные условия принадлежности этой функции пространству $U_{\overline{D}^0}$ выполнены по построению функции $v_{\overline{D}^0}$. Но функция (9) является решением задачи (2), (3) с правой частью (7). Очевидно, что сужение $u_{\overline{D}^0}(x)$ функции (9) на $\overline{D} \subset \overline{D}^0$, т. е. потенциал (8), есть тождественный нуль. □

Теорема 2. *Потенциал $u_{\overline{D}} = P_{\overline{D}\Gamma} v_\Gamma$ при любой плотности $v_\Gamma \in$ $\in U_\Gamma$ есть решение однородного уравнения* (1). *Если $u_{\overline{D}}$ есть решение однородного уравнения* (1) *и допускает доопределение до некоторой $u_{\overline{D}^0} \in U_{\overline{D}^0}$ и если $u_\Gamma = \mathrm{Tr}_{\Gamma\overline{D}} u_{\overline{D}}$ есть данные Коши этого решения, то потенциал $P_{\overline{D}\Gamma} u_\Gamma$ с плотностью u_Γ совпадает с $u_{\overline{D}}$*:

$$u_{\overline{D}} = P_{\overline{D}\Gamma}\, u_\Gamma. \tag{10}$$

Доказательство. Справедливость первого утверждения очевидна, так как решение $u_{\overline{D}^0}$ задачи (2), (3) с правой частью (7) в области D удовлетворяет однородному уравнению (1), поскольку правая часть (7) на D есть нуль. Для доказательства второго утверждения при построении потенциала $u_{\overline{D}} = P_{\overline{D}\Gamma} u_\Gamma$ используем в качестве $v_{\overline{D}^0}$ в (7) ту функцию $v_{\overline{D}^0} \in U_{\overline{D}^0}$, которая является доопределением $u_{\overline{D}}$ на $\overline{D}^0 \backslash D$ до элемента пространства $U_{\overline{D}^0}$. В силу теоремы 1 потенциал допускает такой выбор $v_{\overline{D}^0}$, поскольку

$$\mathrm{Tr}_{\Gamma\overline{D}^0}\, v_{\overline{D}^0} = \mathrm{Tr}_{\Gamma\overline{D}}\, u_{\overline{D}} = u_\Gamma.$$

Прямой подстановкой $v_{\overline{D}^0}$ в уравнения (2), (3) проверяется, что $v_{\overline{D}^0}$ является решением задачи (2), (3) с правой частью (7). Но сужение $v_{\overline{D}^0}$ на $\overline{D} \subset D^0$, т. е. потенциал (8), совпадает с (10). □

2. Проектор Кальдерона. Введем оператор $P_{\Gamma\Gamma}\colon U_\Gamma \to U_\Gamma$, определив его равенством

$$P_{\Gamma\Gamma} = \mathrm{Tr}_{\Gamma\overline{D}} P_{\overline{D}\Gamma}.$$

Теорема 3. *Уравнение*

$$u_\Gamma = P_{\Gamma\Gamma}\, u_\Gamma \tag{11}$$

выполнено для тех и только тех $u_\Gamma \in U_\Gamma$, *которые являются данными Коши для некоторого решения* $u_{\overline{D}}$ *однородного уравнения* (1), *допускающего доопределение всюду в* $\overline{D}^0$ *до некоторой* $u_{\overline{D}^0} \in U_{\overline{D}^0}$. *Оператор* $P_{\Gamma\Gamma}$ *есть проектор*, $P^2_{\Gamma\Gamma} = P_{\Gamma\Gamma}$.

Доказательство. Если u_Γ — данные Коши для некоторого решения однородного уравнения (1), допускающего доопределение до $u_{\overline{D}^0} \in U_{\overline{D}^0}$, то в силу теоремы 2 справедливо (10). Применяя к обеим частям (10) операцию $\mathrm{Tr}_{\Gamma\overline{D}}$, получаем (11). Обратно, пусть $u_\Gamma \in U_\Gamma$ и u_Γ удовлетворяет (11). Рассмотрим потенциал (8), который запишем формулой

$$w_{\overline{D}} = P_{\overline{D}\Gamma} u_\Gamma.$$

Применяя к обеим частям этой формулы операцию $\mathrm{Tr}_{\Gamma\overline{D}}$, получаем $w_\Gamma = P_{\Gamma\Gamma}\, u_\Gamma$. В силу (11) $w_\Gamma = u_\Gamma$. Но $w_{\overline{D}}$ есть решение уравнения (1), так что u_Γ есть данные Коши некоторого решения, а именно решения $w_{\overline{D}}$ уравнения (1).

Докажем, что $P^2_{\Gamma\Gamma} = P_{\Gamma\Gamma}$. Зададим произвольно $v_\Gamma \in U_\Gamma$. Пусть $u_\Gamma = P_{\Gamma\Gamma}\, v_\Gamma$. Применяя к обеим частям равенства $u_\Gamma = P_{\Gamma\Gamma}\, v_\Gamma$ операцию $P_{\Gamma\Gamma}$, получаем $P_{\Gamma\Gamma}\, u_\Gamma = p^2_{\Gamma\Gamma}\, v_\Gamma$. Заметим, что u_Γ есть данные Коши решения $u_D = P_{D\Gamma}\, v_\Gamma$. В силу (11) будет $P_{\Gamma\Gamma}\, u_\Gamma = u_\Gamma$. Поэтому $P_{\Gamma\Gamma}\, u_\Gamma = P^2_{\Gamma\Gamma}\, v_\Gamma$ можно записать так: $u_\Gamma = P^2_{\Gamma\Gamma}\, v_\Gamma$, т. е.

$$P_{\Gamma\Gamma}\, v_\Gamma \underset{v_\Gamma}{\equiv} P^2_{\Gamma\Gamma}\, v_\Gamma. \ \square$$

Обсудим полученный результат. Оказалось, что граничное уравнение с проектором Кальдерона

$$u_\Gamma - P_{\Gamma\Gamma}\, u_\Gamma = 0 \tag{12}$$

равносильно уравнению (1) в том смысле, что оно выполнено для тех и только тех u_Γ, которые являются данными Коши для какого-либо решения задачи (1). Если u_Γ есть решение уравнения (12), то соответствующее решение уравнения (1) восстанавливается по u_Γ по формуле (10).

3. Редукция краевой задачи к равносильным ГУРП. Пусть в дополнение к уравнению (1) задано какое-либо условие

$$l_{\Gamma\Gamma}\, u_\Gamma = \varphi_\Gamma, \quad \varphi_\Gamma \in \Phi_\Gamma, \quad u_\Gamma \in U_\Gamma, \tag{13}$$

где Φ_Γ — некоторое пространство функций на Γ, $l_{\Gamma\Gamma}\colon U_\Gamma \to \Phi_\Gamma$. Тогда пара условий (12), (13) относительно плотности потенциала u_Γ и есть ГУРП, равносильные исходной краевой задаче (1), (13). Очевидно, решение существует и единственно для этих задач одновременно, причем решение задачи (1), (13) выражается через решение задачи (12), (13) по формуле (8).

Представление потенциала в форме (8), а также граничные уравнения с проектором вида (11), полученные на основе использования

потенциала в форме (8), не требуют знания функции Грина и допускают обобщения на широкий класс систем уравнений с переменными коэффициентами относительно вектор-функций многих переменных.

Задачи

1. Пусть $a(x) \equiv b(x) \equiv 1$ и пусть $G(x, y)$ — функция Грина задачи (2), (3).

Показать следующее:

а) потенциал (8) с плотностью $v_\Gamma = \begin{bmatrix} v_0(s) \\ v_1(s) \end{bmatrix}$ можно записать в виде

$$u(x) = \int\limits_\Gamma \left[v_0(y)\frac{\partial G}{\partial n_y} - v_1(y)G\right] ds_y,$$

где s_y — длина дуги вдоль Γ в точке $y \in \Gamma$;

б) условие (11) в точке $x \in \Gamma$ принимает вид

$$v_0(s_x) = \lim_{\substack{x_0 \to x \\ x_0 \in D}} \int\limits_\Gamma \left[v_0(y)\frac{\partial G}{\partial n_y} - v_1(y)G\right] ds_y,$$

$$v_1(s_x) = \lim_{\substack{x_0 \to x \\ x_0 \in D}} \left\{\frac{\partial}{\partial n_x}\int\limits_\Gamma \left[v_0(y)\frac{\partial G}{\partial n_y} - v_1(y)G\right] ds_y\right\}.$$

2. В области $\overline{D}^-$, дополняющей D до всего квадрата $\overline{D}^0$, построить потенциал $u_{\overline{D}^-} = P_{D^-\Gamma} v_\Gamma$, выражающий общее решение в области D^- однородного уравнения (1) и удовлетворяющий нулевым условиям на границе ∂D^0. Построить также граничное уравнение с проектором, аналогичное уравнению (11).

Указание. Принять за новое D область D^- и повторить построения из основного текста.

§ 2. Устойчивость ГУРП и проблема их дискретизации

При редукции краевых задач в области D к ГУРП возникает система уравнений вида

$$u_\Gamma - P_{\Gamma\Gamma}\, u_\Gamma = 0, \quad l_{\Gamma\Gamma}\, u^\Gamma = \varphi_\Gamma, \quad u_\Gamma \in U_\Gamma, \quad \varphi_\Gamma \in \Phi_\Gamma, \qquad (1)$$

где U_Γ, Φ_Γ — некоторые пространства, которые будем считать нормированными.

Мы установим здесь факт устойчивости ГУРП и наметим на его основе подход к их дискретизации для численного решения.

1. Устойчивость ГУРП. Можно показать, что при естественных предположениях об исходной краевой задаче и естественном выборе норм выполнено неравенство

$$\|u_\Gamma\|_{U_\Gamma} \leqslant c\|\varphi_\Gamma\|_{\Phi_\Gamma}, \qquad (2)$$

где c — постоянная, не зависящая от $\varphi_\Gamma \in \Phi_\Gamma$.

Будем считать, что неравенство (2) выполнено. Оно означает устойчивость решения $u_\Gamma \in U_\Gamma$ задачи (1) относительно возмущений $u_\Gamma \in U_\Gamma$ в классе функций u_Γ, для которых равенство $u_\Gamma - P_{\Gamma\Gamma}\, u_\Gamma = 0$ выполнено точно. Можно доказать более сильное утверждение об устойчивости систем вида (1), охватывающее случай $u_\Gamma - P_{\Gamma\Gamma}\, u_\Gamma \neq 0$. Сформулируем его для абстрактных уравнений вида

$$u - Pu = 0, \quad lu = \varphi, \quad u \in U, \quad \varphi \in \Phi, \tag{3}$$

где U, Φ — некоторые линейные нормированные пространства, $P\colon U \to U$, причем $P = P^2$ — проектор;

$$l\colon U \to \Phi,$$

где l — линейный оператор.

Теорема 1. *Пусть для каждой функции $u \in U$, удовлетворяющей уравнению $u - Pu = 0$, выполняется оценка*

$$\|u\|_U \leqslant c_1 \|\varphi\|_\Phi, \tag{4}$$

где c_1 — некоторая постоянная. Пусть, кроме того, операторы P, l ограничены, т. е. $\|P\| < \infty$, $\|l\| < \infty$, где $\|P\| = \sup\limits_u (\|Pu\|_U : \|u\|_U)$; $\|l\| = \sup\limits_u (\|lu\|_\Phi \colon \|u\|_U)$.

Тогда при произвольной $u \in U$ выполнены оценки

$$\|u\|_U^2 \leqslant c_2 (\|u - Pu\|_U^2 + \|lu\|_\Phi^2) \leqslant c_3 \|u\|_U^2, \tag{5}$$

где c_2, c_3 — некоторые постоянные, не зависящие от $u \in U$.

Доказательство. Обозначим $z = u - Pu$, или $u - z = Pu$. Применяя к обеим частям первого равенства оператор P, получаем $Pz = Pu - P^2 u = 0$. Поэтому $u - z = Pu = P(u - z)$. Но из $u - z = P(u - z)$ в силу предположения теоремы следует оценка $\|u - z\|_U \leqslant c_1 \|l(u - z)\|_\Phi$, откуда вытекает

$$\begin{gathered} \|u\|_U \leqslant \|z\|_U + c_1 \|l(u - z)\|_\Phi \leqslant \|z\|_U + c_1 \|lu\|_\Phi + c_1 \|lz\|_\Phi \leqslant \\ \leqslant (1 + c_1 \|l\|)\, \|z\|_U + c_1 \|lu\|_\Phi, \\ \|u\|_U \leqslant (1 + c_1 \|l\|)\, \|u - Pu\|_U + c_1 \|lu\|_\Phi. \end{gathered} \tag{6}$$

Отметим, что для любых вещественных a, b справедливо $2(a^2 + b^2) - (a + b)^2 = (a - b)^2 \geqslant 0$, т. е.

$$(a + b)^2 \leqslant 2a^2 + 2b^2. \tag{7}$$

Положим $a = (1 + c_1 \|l\|)\, \|u - Pu\|_U$, $b = c_1 \|lu\|$, возведем обе части (6) в квадрат и воспользуемся (7). Получим

$$\|u\|_U^2 \leqslant 2(1 + c_1 \|l\|)^2\, \|u - Pu\|^2 + 2c_1^2\, \|lu\|_\Phi^2.$$

Отсюда следует первое из неравенств (5) с постоянной

$$c_2 = 2 \max\, [(1 + c_1 \|l\|)^2, c_1^2].$$

Вторая оценка (5), очевидно, справедлива, если в качестве c_3 принять

$$c_3 = c_2[(1 + \|P\|)^2 + \|l\|^2]. \ \square$$

Отметим важную для дальнейшего теорему.

Теорема 2. *Пусть* $\widetilde{u}$ — *решение системы* (3):

$$\widetilde{u} - P\widetilde{u} = 0, \quad l\widetilde{u} = \varphi, \tag{8}$$

и пусть выполнены условия теоремы 1. *Фиксируем произвольную функцию* $v \in U$. *Отклонение* $\|\widetilde{u} - v\|$ *выбранного* v *от решения* $\widetilde{u}$ *удовлетворяет оценкам*

$$\|\widetilde{u} - v\|_U^2 \leqslant c_2(\|lv - \varphi\|_\Phi^2 + \|v - Pv\|_U^2) \leqslant c_3\|\widetilde{u} - v\|_U^2. \tag{9}$$

Доказательство. В качестве произвольной функции $u \in U$ возьмем $u = \widetilde{u} - v$ и воспользуемся оценками (5). Заметим, что

$$lu = l\widetilde{u} - lv = \varphi - lv, \quad u - Pu = (\widetilde{u} - v) - P(\widetilde{u} - v) = -(v - Pv).$$

Тогда (5) перейдет в (9). □

2. Схема дискретизации ГУРП. Пусть задача (3) при каждом $\varphi \in \Phi$ имеет единственное решение $\widetilde{u} \in U$, причем выполнено неравенство $\|\widetilde{u}\|_U \leqslant c_1\|\varphi\|_\Phi$. Пространство U естественно считать бесконечномерным. Для вычисления приближенного решения необходимо дискретизировать задачу, т. е. описать это приближенное решение с помощью конечного набора чисел, подлежащих отысканию. Мы укажем здесь схему перехода от задачи (3) к задаче, где искомым приближенным решением задачи (3) является элемент $u^{(N)}$ некоторого N-мерного линейного подпространства $U^{(N)}$ пространства U.

Для перехода к N неизвестным числам достаточно записать элементы пространства $U^{(N)}$ в виде выражения, зависящего от N числовых параметров. Это заведомо можно сделать. Например, достаточно выбрать какой-либо базис в $U^{(N)}$ и описывать элементы их координатами в этом базисе.

Итак, пусть при каждом данном натуральном N выбрано $U^{(N)} \subset U$. Введем обозначение

$$K^{(0)}(u) = \|u - Pu\|_U^2 + \|lu - \varphi\|_\Phi^2.$$

Примем за приближенное решение $u^{(N)} \in U^{(N)}$ тот элемент пространства $U^{(N)}$, для которого $K^{(0)}(u)$ принимает наименьшее значение:

$$K^{(0)}(u^{(N)}) \leqslant K^{(0)}(v^{(N)}), \quad v^{(N)} \in U^{(N)}. \tag{10}$$

Таким образом, мы перешли от исходной задачи (3) к задаче об отыскании функции $u^{(N)} \in U^{(N)}$.

Выясним некоторые свойства задачи об отыскании $u^{(N)}$. Обозначим

$$\varepsilon_N = \min_{v \in U^{(N)}} \|\widetilde{u} - v\|_U. \tag{11}$$

Теорема 3. *Погрешность* $\widetilde{u} - u^{(N)}$ *приближения* $u^{(N)}$ *удовлетворяет оценке*

$$\|\widetilde{u} - u^{(N)}\|_U^2 \leqslant c_3 \varepsilon_N^2, \tag{12}$$

где c_3 — *постоянная из теоремы* 2.

Доказательство. Пусть $v^{(N)} \in U^{(N)}$ — элемент, для которого достигается минимум в правой части (11), т. е.

$$\varepsilon_N = \|\widetilde{u} - v^{(N)}\|_U.$$

В силу второго неравенства (9) справедливо

$$c_2 K^{(0)}(v^{(N)}) \leqslant c_3 \|\widetilde{u} - v^{(N)}\|_U^2 = c_3 \varepsilon^2.$$

Учитывая (10), можно написать

$$c_2 K^{(0)}(u^{(N)}) \leqslant c_2 K^{(0)}(v^{(N)}) \leqslant c_3 \varepsilon_N^2. \tag{13}$$

Но в силу первого неравенства (9), т. е. неравенства $\|\widetilde{u} - u^{(N)}\|_U \leqslant c_2 K^{(0)}(u^{(N)})$, из (13) следует (12). □

Следствие. *Если последовательность пространств* $U^{(N)}$ ($N = N_1, N_2, \ldots$) *выбрана так, что последовательность чисел* $\varepsilon_1, \varepsilon_2, \ldots$ *стремится к нулю, то в силу* (12) *последовательность приближений* $u^{(N)}$ *сходится к решению задачи* (3).

Для отыскания элемента $u^{(N)}$, придающего наименьшее значение функционалу $K^{(0)}$ на подпространстве $U^{(N)}$, выберем базис $\psi_1^{(N)}$, $\psi_2^{(N)}, \ldots, \psi_N^{(N)}$ в этом подпространстве и запишем искомый элемент $u^{(N)}$ в виде

$$u^{(N)} = \sum_{k=1}^{N} c_k \psi_k^{(N)}, \quad c_k = c_k(N), \quad k = 1, 2, \ldots, N, \tag{14}$$

где c_k — коэффициенты разложения $u^{(N)}$ по выбранному базису.

Будем считать в дальнейшем, что пространства U, Φ суть пространства с некоторыми скалярными произведениями $(u, v)_U$, $(\varphi, \psi)_\Phi$ и нормами $\|u\|_U^2 = (u, u)_V$, $\|\varphi\|_\Phi^2 = (\varphi, \varphi)_\Phi$.

Тогда

$$K^{(0)}(u^{(N)}) = \Big(\sum_{k=1}^{N} c_k (\psi_k^{(N)} - P\psi_k^{(N)}), \quad \sum_{k=1}^{N} c_k (\psi_k^{(N)} - P\psi_k^{(N)}) \Big)_U + \\ + \Big(\sum_{k=1}^{N} c_k l \psi_k^{(N)} - \varphi, \quad \sum_{k=1}^{N} c_k l_k \psi_k^{(N)} - \varphi \Big)_\Phi, \tag{15}$$

т. е. $K^{(0)}(u^{(N)})$ есть квадратичная функция от N чисел $c_1, c_2, \ldots, c_N$ строго положительная в случае $\varphi = 0$, $c_j \not\equiv 0$.

Отыскание элемента (14), придающего минимум выражению (15), равносильно вычислению решения $c = \{c_k\}$ системы линейных уравнений

$$\frac{\partial K^{(0)}(u^{(N)})}{\partial c_k} = 0, \quad k = 1, 2, ..., N,$$

которую можно записать в виде

$$C^{(N)}\mathbf{c} = \mathbf{f}. \tag{16}$$

Здесь $C^{(N)} = (c_{ij}^{(N)})$ $(i, j = 1, 2, ..., N)$ — некоторая матрица, а $\mathbf{f}$ — некоторый N-мерный вектор. Для вычисления элементов $\{c_{ij}^{(N)}\}$ и компонент вектора $\mathbf{f}$ нужно, очевидно, уметь вычислять элементы $P\psi_k^{(N)} \in U$ и $l\psi_k^{(N)} \in \Phi$ $(k = 1, 2, ..., N)$ и соответствующие скалярные произведения вида

$$(P\psi_i^{(N)}, \psi_j^{(N)}), \quad (P\psi_i^{(N)}, P\psi_j^{(N)})_U, \quad (l\psi_i^{(N)}, l\psi_j^{(N)})_\Phi, \\ i, j = 1, 2, ..., N. \tag{17}$$

Затем остается точно или методом итераций (приближенно) решить систему (16) и получить искомое решение (14).

Замечание 1. Переход к системе (16) в канонической записи, требующий вычисления матрицы $C^{(N)}$, не является единственным возможным путем для вычисления элемента (14), минимизирующего $K^{(0)}(v^{(N)})$. Часто это и не самый дешевый по затратам ресурсов компьютера (времени и памяти) путь. Однако будем придерживаться только этого пути из-за того, что тогда удастся осуществить изложение более легко.

Альтернативный подход отыскания элемента $u^{(N)} \in U^N$, минимизирующего выражение $K^{(0)}(v^{(N)})$, состоит в использовании операторной записи для СЛАУ, равносильной задаче об отыскании минимизирующего элемента для $K^{(0)}(v^{(N)})$ $(v^{(N)} \in U^N)$ с последующим решением этой СЛАУ итерациями, на каждом шаге которых требуется вычислять результат действия оператора СЛАУ на предыдущее приближение. Этот подход изложен в [16]. Здесь мы о нем говорить больше не будем.

При получении выражений (17), а также матрицы $C^{(N)}$ и правой части $\mathbf{f}$ системы (16) достаточно уметь находить интересующие нас величины лишь приближенно. Однако вычислительный алгоритм должен зависеть от некоторого параметра (обозначим этот параметр h $(h > 0)$) таким образом, чтобы при $h \to 0$ результаты вычислений становились сколь угодно точными. Это позволит выбрать и фиксировать достаточно малое $h > 0$ так, чтобы получить требуемую точность.

Замечание 2. Пусть с целью получения вычислительных формул выражения $K^{(0)}(v^{(N)})$, Pv заменены приближенными выражениями $K_h^{(0)}(v^{(N)})$, $P_h v^{(N)}$, уточняющимися при $h \to 0$. При этом с уменьшением h возрастают вычислительные затраты. При заданном $\varepsilon > 0$ наибольшее число $h > 0$, при котором выполнены неравенства

$$\begin{aligned} &\left| K^{(0)} \frac{v^{(N)}}{\|v^{(N)}\|} - K_h^{(0)} \frac{v^{(N)}}{\|v^{(N)}\|} \right| \leqslant \varepsilon, \\ &\frac{1}{\|v^{(N)}\|} \|Pv^{(N)} - P_h v^{(N)}\|_U < \varepsilon, \end{aligned} \tag{18}$$

вообще говоря, зависит не только от h, но и от $v^{(N)}$, $h = h(\varepsilon, v^{(N)})$.

Обозначим $h_{\min}(\varepsilon) = \min\limits_{v^{(N)} \in U^{(N)}} h(\varepsilon, v^{(N)})$. Может оказаться, что на элементах $v^{(N)}$, близких к решению $u^{(N)}$, для выполнения неравенств (18) допускается выбор числа $h(\varepsilon, v^{(N)})$, гораздо менее жесткий, чем $h_{\min}(\varepsilon)$; это справедливо, если

$$h(\varepsilon, u^{(N)}) \colon h_{\min}(\varepsilon) \gg 1. \tag{19}$$

При отыскании $u^{(N)}$ в процессе итераций возникают $v^{(N)}$, близкие к $u^{(N)}$, так что достаточно использовать именно $h(\varepsilon, u^{(N)})$, а не $h_{\min}(\varepsilon, v^{(N)})$. Вычисления будут более дешевыми, чем было бы при использовании $h_{\min}$.

Соотношение вида (19), усиливающееся с ростом N, как мы увидим, имеет место при приложении общей схемы к краевым задачам для уравнения Лапласа. Благодаря этому окажется, что точность вычисления элементов c_{ij} матрицы $C^{(N)}$ может быть меньше, чем точность отыскания $u^{(N)}$.

3. Задача Дирихле для уравнения Лапласа и подходящие нормы в пространствах U_Γ, Φ_Γ. Вернемся от абстрактной задачи (3) с проектором $P \colon U \to U$ к ГУРП вида (1) для задачи

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = 0, \quad x \in D; \quad u_D\big|_\Gamma = \varphi(s). \tag{20}$$

Эта задача Дирихле для уравнения Лапласа была сведена к ГУРП (12), (13) из § 1

$$u_\Gamma - P_{\Gamma\Gamma}\, u_\Gamma = 0_\Gamma, \quad u_{\bar D}\big|_\Gamma = \varphi(s).$$

Будем считать длину Γ равной единице. Отнесем к U_Γ и к Φ_Γ “достаточно гладкие” вектор-функции $v_\Gamma = \begin{bmatrix} v^{(0)}(s) \\ v^{(1)}(s) \end{bmatrix}$ и “достаточно гладкие” функции $\varphi_\Gamma = \varphi(s)$ соответственно.

Замечание. “Достаточно гладкими” будем считать $v^{(0)}(s)$, $v^{(1)}(s)$, $\varphi(s)$, если они принадлежат соответственно пространствам

Гёльдера

$$v^{(0)} \in C_{k+2+\alpha,\Gamma}, \quad v^{(1)} \in C_{k+1+\alpha,\Gamma}, \quad \varphi(s) \in C_{k+2+\alpha,\Gamma},$$

где $k \geqslant 0$ и $0 < \alpha < 1$ — фиксированные числа.

Пространством Гёльдера $C_{j+\alpha,\Gamma}$ называется пространство периодических с периодом 1 функций $f(s)$, имеющих производные по s до порядка j, которые удовлетворяют условию Гёльдера с показателем α:

$$\left|\frac{d^j f(s'')}{ds^j} - \frac{d^j f(s')}{ds^j}\right| \leqslant \mathrm{const} \cdot |s'' - s'|^{\alpha}.$$

Мы описали запас элементов, входящих в пространства U_Γ, Φ_Γ. Введем теперь нормы, положив

$$\begin{aligned} \|v_\Gamma\|^2_{U_\Gamma} &= \int\limits_\Gamma \left[|v^{(0)}(s)|^2 + \left|\frac{dv^{(0)}}{ds}\right|^2 + |v^{(1)}(s)|^2\right] ds, \\ \|\varphi_\Gamma\|^2_{\Phi_\Gamma} &= \int\limits_\Gamma \left[|\varphi(s)|^2 + \left|\frac{d\varphi}{ds}\right|^2\right] ds. \end{aligned} \qquad (21)$$

Возникающие при этом нормированные пространства U_Γ, Φ_Γ не являются так называемыми полными пространствами, но это нам и не нужно.

Введенные нормы (21) можно считать евклидовыми, если ввести скалярные произведения в U_Γ и Φ_Γ формулами

$$(u_\Gamma, v_\Gamma)_{U_\Gamma} = \int\limits_0^1 \left[u^{(0)} v^{(0)} + \frac{du^{(0)}}{ds}\frac{dv^{(0)}}{ds} + u^{(1)} v^{(1)}\right] ds, \qquad (22)$$

$$(\varphi, \psi)_{\Phi_\Gamma} = \int\limits_0^1 \left[\varphi(s)\psi(s) + \frac{d\varphi}{ds}\frac{d\psi}{ds}\right] dt. \qquad (23)$$

Из теории эллиптических уравнений следует, что оператор $P_{\Gamma\Gamma}$: $U_\Gamma \to U_\Gamma$ ограничен. Кроме того, решение задачи (20) удовлетворяет оценке

$$\|u_\Gamma\|_{U_\Gamma} \leqslant \mathrm{const} \cdot \|\varphi_\Gamma\|_{\Phi_\Gamma}. \qquad (24)$$

Таким образом, в рассматриваемом случае реализуются условия, предусмотренные при общих построениях для абстрактной задачи с проектором (3).

Для дискретизации задачи в соответствии с общей схемой сначала надо выделить подпространство $U_\Gamma^{(N)} \subset U_\Gamma$. Сделаем это следующим образом. Зададим $H > 0$ так, чтобы число $1/H$ оказалось целым числом, и положим $N = 2H^{-1}$. Отметим на границе Γ (длина которой равна единице) равноотстоящие точки $s_j = jH$ $(j = 1, 2, ..., n)$; $n = H^{-1}$.

Подпространство $U^{(N)}$ определим как пространство вектор-функции $v_\Gamma^{(N)} = \begin{bmatrix} v^{(0,N)} \\ v^{(1,N)} \end{bmatrix}$ с компонентами $v^{(0,N)}(s)$, $v^{(1,N)}(s)$ вида

$$v^{(0,N)}(s) = \sum_{j=1}^{n} v_j^{(0)} \psi^{(0,j)}(s), \tag{25}$$

$$v^{(1,N)}(s) = \sum_{j=1}^{n} v_j^{(1)} \psi^{(1,j)}(s), \tag{26}$$

где $v_j^{(0)}$, $v_j^{(1)}$ $(j = 1, 2, ..., n)$ — произвольные числа, а

$$\psi^{(0,j)}(s) \in C_{k+2+\alpha,\Gamma}, \quad j = 1, 2, ..., n,$$

$$\psi^{(1,j)}(s) \in C_{k+1+\alpha,\Gamma}, \quad j = 1, 2, ..., n,$$

— кусочно многочленные базисные функции. Они подобраны так, чтобы выражения (25) и (26) были локальными гладкими сплайнами, имеющими соответственно $k+2$ и $k+1$ непрерывных производных, принимающих в точке s_j значения $v_j^{(0)}$ и $v_j^{(1)}$ $(j = 1, ..., n)$ соответственно. Такие сплайн-функции описаны в § 3 гл. 1. Они будут использованы нами лишь при теоретических рассмотрениях, а для вычислений можно в качестве функций (25), (26) использовать обычную кусочно многочленную интерполяцию.

Обозначим через $\Psi_\Gamma^{(N,j)} \in U_\Gamma$ вектор-функции

$$\Psi_\Gamma^{(N,j)} = \begin{bmatrix} \psi^{(0,j)}(s) \\ 0 \end{bmatrix}, \quad j = 1, 2, ..., n,$$

$$\Psi_\Gamma^{(N,j)} = \begin{bmatrix} 0 \\ \psi^{(1,j-n)}(s) \end{bmatrix}, \quad j = n+1, n+2, ..., N.$$

Тогда элемент $v_\Gamma \in U_\Gamma^{(N)}$ можно будет записать в виде (14):

$$v_\Gamma = \sum_{j=1}^{N} v_j \Psi_\Gamma^{(N,j)}. \tag{27}$$

Обозначим через $P_\Gamma^{(0)} \Psi_\Gamma^{(N,j)}$ и $P_\Gamma^{(1)} \Psi_\Gamma^{(N,j)}$ соответственно первую и вторую функции из составляющих вектор-функцию $P_\Gamma \Psi_\Gamma^{(N,j)}$, $P_\Gamma \equiv \equiv P_{\Gamma\Gamma}$.

Скалярные произведения (17), используемые для построения матрицы $C^{(N)}$ системы (16), в данном случае примут вид

$$(P_\Gamma \Psi_\Gamma^{(N,i)}, \Psi_\Gamma^{(N,j)})_{U_\Gamma}, \quad (P_\Gamma \Psi_\Gamma^{(N,i)}, P_\Gamma \Psi_\Gamma^{(N,j)})_{U_\Gamma}, (l\Psi_\Gamma^{(N,i)}, l\Psi_\Gamma^{(N,j)})_{\Phi_\Gamma}. \tag{28}$$

Обозначим первую и вторую компоненты вектор-функции $P_{\Gamma\Gamma} \Psi_\Gamma^{(N,j)} \in \in U_\Gamma$ через $P_{\Gamma\Gamma}^{(0)} \Psi_\Gamma^{(N,j)}$ и $P_{\Gamma\Gamma}^{(1)} \Psi_\Gamma^{(N,j)}$ соответственно. В этом случае чис-

ла (28), очевидно, будут выражены через следующие интегралы:

$$\int_0^1 P_{\Gamma\Gamma}^{(0)} \Psi_\Gamma^{(N,i)} \psi^{(0,j)} ds, \qquad i = 1, 2, ..., N, \quad j = 0, 1, ..., n, \tag{29}$$

$$\int_0^1 \frac{d}{ds}(P_{\Gamma\Gamma}^{(0)} \Psi_\Gamma^{(N,i)}) \frac{d}{ds}\psi^{(0,j)} ds, \qquad i = 1, 2, ..., N, \quad j = 0, 1, ..., n, \tag{30}$$

$$\int_0^1 P_{\Gamma\Gamma}^{(1)} \Psi_\Gamma^{(N,i)} \psi^{(1,j)} ds, \qquad i = 1, 2, ..., N, \quad j = 1, 2, ..., n, \tag{31}$$

$$\int_0^1 (P_{\Gamma\Gamma}^{(0)} \Psi_\Gamma^{(N,i)})(P_{\Gamma\Gamma}^{(0)} \Psi_\Gamma^{(N,j)})\, ds, \qquad i, j = 1, 2, ..., N, \tag{32}$$

$$\int_0^1 \frac{d}{ds}(P_{\Gamma\Gamma}^{(0)} \Psi_\Gamma^{(N,i)}) \frac{d}{ds}(P_{\Gamma\Gamma}^{(0)} \Psi_\Gamma^{(N,j)})\, ds, \qquad i, j = 1, 2, ..., N, \tag{33}$$

$$\int_0^1 (P_{\Gamma\Gamma}^{(1)} \Psi_\Gamma^{(N,i)})(P_{\Gamma\Gamma}^{(1)} \Psi_\Gamma^{(N,j)})\, ds, \qquad i, j = 1, 2, ..., N, \tag{34}$$

$$\int_0^1 \psi^{(0,i)} \psi^{(1,j)} ds, \qquad i, j = 1, 2, ..., n. \tag{35}$$

Интегралы (29)–(35) можно приближенно заменить суммами с помощью квадратурных формул. Именно, выбрав некоторое натуральное N' и узлы $s_j = j/N'$ $(j = 0, 1, ..., N')$, можно воспользоваться формулами трапеций.

Можно добиться сколь угодно большой точности, выбирая число N' достаточно большим. При вычислении выражений (35) проблем нет. Однако слагаемые в возникающих при замене интегралов(29)–(34) суммах выражаются через числа

$$\begin{gathered} P_\Gamma^{(0)} \Psi_\Gamma^{(N,j)}\big|_{s=s_k}, \quad \frac{d}{ds} P_\Gamma^{(0)} \Psi_\Gamma^{(N,j)}\big|_{s=s_k}, \quad P_\Gamma^{(1)} \Psi_\Gamma^{(N,j)}\big|_{s=s_k}. \\ j = 1, 2, ..., N, \quad k = 1, 2, ..., N', \end{gathered} \tag{36}$$

Для отыскания чисел (36) мы воспользуемся аппроксимацией потенциала Кальдерона $P_{D\Gamma} v_\Gamma$ с заданной плотностью $v_\Gamma = \begin{bmatrix} v^{(0)}(s) \\ v^{(1)}(s) \end{bmatrix}$ разностными потенциалами, которые эффективно вычисляются.

Следующий параграф (§ 3) посвятим определению, изучению свойств и описанию алгоритма вычисления разностных потенциалов. Наконец, в § 4 будет рассказано об использовании разностных потенциалов для приближенного отыскания чисел (36), а также о некоторых других приложениях метода разностных потенциалов.

§ 3. Определение и вычисление разностных потенциалов

Зададим натуральное число K и положим $h = 1/K$. На плоскости xOy введем сетку $(x_{m_1}, y_{m_2}) = (m_1 h, m_2 h)$ $(m_1, m_2 = 0, \pm 1, ...)$ и рассмотрим пятиточечный разностный аналог уравнения Пуассона

$$\sum_{n \in N_m} a_{mn} u_n = f_m,$$

где N_m — шаблон разностной схемы в точке (x_{m_1}, y_{m_2}), а именно, совокупность пяти точек $(x_{m_1} \pm h, y_{m_2} \pm h)$, (x_{m_1}, y_{m_2}). Обозначим через u_n значение искомой сеточной функции в точке сетки $(x_{n_1}, y_{n_2}) = (n_1 h, n_2 h)$, через f_m заданное значение сеточной функции в точке $(x_{m_1}, y_{m_2}) = (m_1 h, m_2 h)$, а через a_{mn} коэффициенты:

$$a_{mn} = \begin{cases} h^{-2}, & n_1 = m_1 \pm 1 \quad \text{или} \quad n_2 = m_2 \pm 1, \\ -4h^{-2}, & n = m \quad (n_1 = m_1, n_2 = m_2). \end{cases}$$

Пусть $D = D^+$ — некоторая ограниченная область с границей Г. Пусть для определенности она вместе со своей границей лежит внутри квадрата $D^0 = \{0 < x < 1,\ 0 < y < 1\}$. Обозначим через M^0 множество точек сетки, попавших внутрь квадрата D^0, через M^+ множество точек сетки, попавших строго внутрь области D. Определим M^- как дополнение множества M^+ до M^0, т. е. положим $M^- = M^0 \backslash M^+$. Введем множества N^0, N^+, N^- и сеточную границу γ сеточных областей N^+, N^-. Положим $N^0 = \cup N_m$ $(m \in M^0)$, т. е. N^0 — это то множество, которое заметает пятиточечный шаблон N_m, когда m пробегает множество M^0. Аналогично, $N^+ = \cup N_m$ $(m \in M^+)$, $N^- = \cup N_m$ $(m \in M^-)$. Наконец, определим границу γ, положив $\gamma = N^+ \cap N^-$. (Здесь $\cup$ и $\cap$ — знаки объединения и пересечения множеств.) Элементы границы γ на рис. 42 изображены точками.

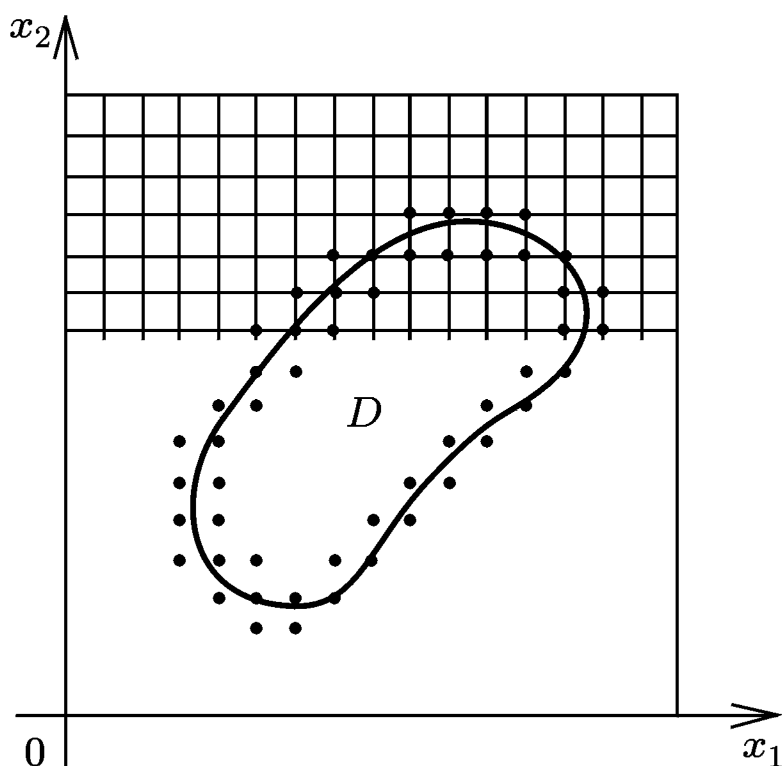


Рис. 42

Введем разностную вспомогательную задачу относительно функции u_{N^0}, определенную на N^0:

$$\sum_{n \in N_m} a_{mn} u_n = f_m, \quad m \in M^0, \quad u_n\big|_{\partial D^0} = 0, \tag{1}$$

где f_m — произвольная функция, определенная на M^0.

Отметим, что задача (1) всегда имеет одно и только одно решение, причем это решение можно вычислить методом Фурье (методом разделения переменных).

Введем пространство V_γ всех функций, определенных на γ.

Определение. *Разностным потенциалом* $u_{N^+} = P^+_{N^+\gamma} v_\gamma$ *с плотностью* v_γ называется функция u_{N^+}, определенная на множестве N^+ и совпадающая на множестве N^+ ($N^+ \subset N^0$) с решением z_{N^0} следующей задачи вида (1):

$$\sum_{n\in N_m} a_{mn} z_n = \begin{cases} 0, & m \in M^+, \\ \sum\limits_{n\in N_m} a_{mn} v_n, & m \in M^-, \end{cases} \tag{2}$$

$$z_n\big|_{\partial D^0} = 0, \tag{3}$$

где v_n — произвольная функция, которая определена на N^0, обращается в нуль в точках, лежащих на сторонах квадрата D^0, т. е. на ∂D^0, и совпадает с заданной плотностью v_γ в точках $n \in \gamma$.

Отметим сразу же, что функция z_n $(n \in N^0)$ может быть легко вычислена как решение задачи (2), (3), являющейся частным случаем задачи вида (1).

Теорема 1. *Определение разностного потенциала* $u_{N^+} = P^+_{N^+\gamma} v_\gamma$ *корректно в том смысле, что* u_{N^+} *зависит только от* v_γ, *но не от способа доопределения функции* v_γ *до некоторой функции* v_{N^0} *в точках* $N^0 \backslash \gamma$.

Доказательство. Разобьем v_{N^0} на два слагаемых: $v_{N^0} = v^{\mathrm{I}}_{N^0} + v^{\mathrm{II}}_{N^0}$, так что

$$v^{\mathrm{I}}_{N^0} = \begin{cases} 0, & n \notin \gamma, \\ v_\gamma\big|_n, & n \in \gamma, \end{cases} \qquad v^{\mathrm{II}}_{N^0} = \begin{cases} v_{N^0}\big|_n, & n \notin \gamma, \\ 0, & n \in \gamma. \end{cases}$$

В силу линейности задачи (2), (3) для доказательства теоремы достаточно показать, что решение $u^{\mathrm{II}}_{N^0}$ задачи

$$\sum_{n\in N_m} a_{mn} u^{\mathrm{II}}_n = \begin{cases} 0, & m \in M^+, \\ \sum\limits_{n\in N_m} a_{mn} v^{\mathrm{II}}_n, & m \in M^-, \end{cases} \tag{4}$$

$$u_n\big|_{\partial D^0} = 0$$

обращается в нуль на N^+.

Для этого покажем, что решение задачи (4), (5) задается формулой

$$u^{\mathrm{II}}_n = \begin{cases} 0, & n \in N^+, \\ v^{\mathrm{II}}_n, & n \in N^-, \end{cases} \tag{5}$$

и тем самым обращается в нуль на M^+. Во-первых, заметим, что функция u_n^{II}, задаваемая формулой (5), однозначна, так как в случае $n \in \gamma = N^+ \cap N^-$ обе строчки формулы (6) принимают одно и то же (нулевое) значение. Очевидно, что u_n^{II} удовлетворяет уравнениям (4). □

Теорема 2. *Разностный потенциал $u_{N^+} = P^+_{N^+\gamma} u_\gamma$ при произвольной плотности v_γ удовлетворяет однородному уравнению*

$$\sum_{n \in N_m} a_{mn} u_n = 0, \quad m \in M^+. \tag{6}$$

Доказательство. Справедливость теоремы непосредственно следует из определения разностного потенциала, а именно, из равенства (2) при $m \in M^+$. □

Теорема 3. *Пусть функция v_γ допускает такое доопределение v_{N^+} всюду на N^+, при котором функция v_n $(n \in N^+)$ удовлетворяет однородному уравнению* (6). *В этом случае такое доопределение единственно, и разностный потенциал $u_{N^+} = P^+_{N^+\gamma} v_\gamma$ есть функция, совпадающая с решением v_{N^+} уравнения* (6).

Доказательство. Пусть условие теоремы выполнено. Доопределим v_γ всюду на N^+ до решения уравнения (6), в точках $n \in \partial D^0$ положим $u_n = 0$, а в остальных точках множества $n \in N^0$ зададим v_n произвольно. Уравнение (2), лежащее в основе определения разностного потенциала, в таком случае можно переписать так:

$$\sum_{n \in N_m} a_{mn} z_n = \sum_{n \in N_m} a_{mn} v_n, \quad m \in M^0, \tag{7}$$

поскольку справедливо

$$\sum_{n \in N_m} a_{mn} v_n \equiv 0, \quad m \in M^+,$$

по построению функции v_n $(n \in N^0)$.

Но уравнение (7) при условии (3) имеет решение $z_n \equiv v_n$ $(n \in N^0)$. В частности, если $n \in N^+$, то в силу определения $u_{N^+} = z_{N^+}$, а значит, $u_{N^+} = v_{N^+}$, и (6) справедливо. □

Теорема 4. *Функция $u_\gamma \in V_\gamma$ может быть доопределена всюду на N^+ до некоторого решения u_{N^+} уравнения* (6), *если выполнено равенство*

$$P^+_{\gamma\gamma} v_\gamma = v_\gamma, \tag{8}$$

и только в этом случае. Здесь $P^+_{\gamma\gamma} v_\gamma$ есть сужение на множество γ разностного потенциала $P^+_{N^+\gamma} v_\gamma$, который определен всюду на N^+, и, в частности, в точках множества $\gamma = N^+ \cap N^-$, которое содержится в N^+; $P^+_{\gamma\gamma}\colon V_\gamma \to V_\gamma$.

Доказательство. Если функция v_γ удовлетворяет условию (8), то она совпадает с сужением на γ разностного потенциала $P^+_{N^+\gamma}v_\gamma$, т. е. этот разностный потенциал представляет собой ее доопределение всюду на N^+. Но разностный потенциал $u_{N^+} = P^+_{N^+\gamma}v_\gamma$ является решением однородного уравнения (6), и первое утверждение теоремы доказано. Обратно, если равенство (8) не выполнено, то функцию v_γ нельзя доопределить до решения v_{N^+} уравнения (6).

В противном случае в силу теоремы 3 разностный потенциал $u_{N^+}= P^+_{N^+\gamma}v_\gamma$ совпадал бы с этим доопределением v_{N^+}, а его сужение $P^+_{\gamma\gamma}v_\gamma$ на γ совпадало бы с сужением v_γ функции v_{N^+} на множество γ, т. е. выполнялось бы (8). □

Аналогично разностному потенциалу $u_{N^+} = P^+_{N^+\gamma}v_\gamma$ с плотностью v_γ, определенному на N^+, можно построить разностный потенциал $u_{N^-} = P^-_{N^-\gamma}v_\gamma$, определенный на N^-.

Определение. *Разностным потенциалом* $u_{N^-} = P^-_{N^-\gamma}v_\gamma$ *с плотностью* v_γ назовем функцию u_{N^-}, определенную на множестве N^- и совпадающую на этом множестве N^- ($N^- \subset N^0$) с решением w_{N^0} следующей задачи вида (1):

$$\sum_{n\in N_m} a_{mn}w_n = \begin{cases} \sum\limits_{n\in N_m} a_{mn}v_n, & m \in M^+, \\ 0, \quad m \in M^-, \end{cases} \tag{9}$$

$$w_n\big|_{\partial D^0} = 0, \tag{10}$$

где v_n — произвольная функция, которая определена на N^0, совпадает с v_γ на $\gamma \subset N^0$ и удовлетворяет условию 3.

Справедливы теоремы об $u_{N^-} = P^-_{N^-\gamma}v_\gamma$, аналогичные теоремам об $u_{N^+} = P^+_{N^+\gamma}v_\gamma$. Мы сформулируем эти теоремы, оставляя доказательство их в качестве упражнений читателям.

Теорема 1′. *Определение разностного потенциала* $u_{N^-} = P^-_{N^-\gamma}v_\gamma$ *корректно в том смысле, что* u_{N^-} *зависит только от* v_γ*, но не от выбранного доопределения функции* v_γ *до функции* v_{N^0}*, использованной в уравнении* (9).

Теорема 2′. *Разностный потенциал* $u_{N^-} = P^-_{N^-\gamma}v_\gamma$ *при произвольной плотности* v_γ *удовлетворяет однородному уравнению*

$$\sum_{n\in N_m} a_{mn}u_n = 0, \quad m \in M^-. \tag{11}$$

Теорема 3′. *Пусть функция* v_γ *допускает такое доопределение* v_n ($n \in N^-$) *всюду на* N^-*, при котором функция* v_n ($n \in N^-$) *удовлетворяет условию* (10) *и однородному уравнению*

$$\sum_{n\in N_m} a_{mn}v_n \equiv 0, \quad m \in M^-. \tag{12}$$

Тогда это доопределение единственно и разностный потенциал $u_{N^-} = P^-_{N-\gamma}v_\gamma$ *есть функция, совпадающая с этим доопределением* v_{N^-}:

$$v_{N^-} = P^-_{N-\gamma}v_\gamma.$$

Теорема 4$'$. *Функция* v_γ *может быть доопределена всюду на* N^- *до решения уравнения* (12) *при условии* (10) *в том и только том случае, если выполнено равенство*

$$P^-_{\gamma\gamma}v_\gamma = v_\gamma. \tag{13}$$

Здесь $P^-_{\gamma\gamma}v_\gamma$ — *сужение на* γ *разностного потенциала* $P^-_{N-\gamma}v_\gamma$; $P^-_{\gamma\gamma}\colon V_\gamma \to V_\gamma$.

Существует связь между потенциалами $u_{N^+} = P^+_{N^+\gamma}v_\gamma$ и $u_{N^-} = P^-_{N-\gamma}v_\gamma$. Эта связь выражается, в частности, теоремой 5. Разобьем границу γ на два подмножества γ^+, γ^-: $\gamma = \gamma^+ \cup \gamma^-$, положив

$$\gamma^+ = \gamma \cap M^+, \quad \gamma^- = \gamma \cap M^-. \tag{14}$$

Теорема 5. *Условие* (8) *выполняется в том и только том случае, если решение* w_n $(n \in N^0)$ *задачи* (9), (10) *удовлетворяет условию*

$$w_n = 0, \quad n \in \gamma^+. \tag{15}$$

Условие (13) *выполняется в том и только том случае, если решение* z_n $(n \in N^0)$ *задачи* (2), (3) *удовлетворяет условию*

$$z_n = 0, \quad n \in \gamma^-. \tag{16}$$

Доказательство. Докажем первое утверждение. Прежде всего заметим, что функция $w_{N^-} = \{w_n\}$ $(n \in N^-)$ в точках $n \in M^-$ удовлетворяет в силу (9) разностному аналогу уравнения Лапласа. Поэтому в силу принципа максимума наибольшее на N^- значение модуля w_n достигается хотя бы в одной точке множества N^-, не принадлежащей M^-, т. е. хотя бы в одной точке $N^-\backslash M^-$. Но множество $N^-\backslash M^-$ состоит из части точек множества γ^+, а также из точек сетки, лежащих на границе ∂D^0 квадрата D^0. В точках, принадлежащих ∂D^0, функция w_n обращается в нуль в силу (10). Поэтому условие (15) необходимо и достаточно для того, чтобы функция w_n тождественно обращалась в нуль всюду на N^-. Теперь очевидно, что условие (15) равносильно также требованию

$$w_n \equiv 0, \quad n \in \gamma. \tag{17}$$

Очевидно, что функция $v_{N^0} = \{v_n\}$ удовлетворяет тождеству

$$\sum_{n \in N_m} a_{mn}v_n = \sum_{n \in N_m} a_{mn}v_n, \quad m \in M^0, \tag{18}$$

$$v_n\Big|_{\partial D^0} = 0. \tag{19}$$

Вычитая из тождества (18) уравнение (2) почленно и обозначая $w_n = v_n - z_n$ $(n \in N^0)$, получаем, очевидно, что w_n является решением задачи (9), (10). В частности, для $n \in N^+$ функция z_n совпадает со значением в этой точке потенциала $u_{N^+} = P^+_{N^+\gamma} v_\gamma$. Поэтому можно написать

$$w_{N^+} = v_{N^+} - P^+_{N^+\gamma} v_\gamma. \tag{20}$$

В частности, на множестве $\gamma \subset N^+$ оно превращается в равенство

$$w_\gamma = v_\gamma - P^+_{\gamma\gamma} v_\gamma,$$

а равенство (8) в равенство $w_\gamma = 0$, или, что то же самое, в равенство (17). □

Мы предоставляем читателю доказать в качестве упражнения второе утверждение теоремы самостоятельно.

Существует глубокая аналогия между интегралами Коши и типа Коши

$$f(z) = \frac{1}{2\pi i} \oint\limits_{\Gamma} \frac{\varphi(\xi)}{\xi - z}\, d\xi \tag{21}$$

из теории аналитических функций и разностными потенциалами (а также континуальными потенциалами $u_D = P_{D\Gamma} v_\Gamma$).

Доопределим потенциал $u_{N^+} = P^+_{N^+\gamma} v_\gamma$ всюду на N^0, положив

$$u^+_{N^0} = P^+_{N^0\gamma} v_\gamma = \begin{cases} P^+_{N^+\gamma} v_\gamma & \text{на} \quad N^+, \\ -(P^-_{N^-\gamma} v_\gamma) & \text{на} \quad N^-. \end{cases} \tag{22}$$

Функция $u^+_{N^0}$, очевидно, двузначна на множестве $\gamma = N^+ \cap N^-$.

Функция (22) и есть разностный аналог интеграла типа Коши (21) (с обходом контура $\Gamma = \partial D$ против часовой стрелки). При этом роль контура Γ играет сеточная граница γ. Роль $\varphi(\xi)$ $(\xi \in \Gamma)$ играет v_γ. Роль аналитических функций играют решения разностного аналога уравнения Лапласа.

Будем называть сеточные функции

$$u^{(\mathrm{in})}_\gamma = P^+_{\gamma\gamma} v_\gamma \quad \text{и} \quad u^{(\mathrm{ex})}_\gamma = -P^-_{\gamma\gamma} v_\gamma$$

предельными значениями функции (22) соответственно *изнутри и извне сеточной области* N^+.

Читатель легко проверит, что сеточная функция

$$P^+_{\gamma\gamma} v_\gamma - (-P^-_{\gamma\gamma} v_\gamma) = (P^+_{\gamma\gamma} + P^-_{\gamma\gamma}) v_\gamma$$

совпадает с плотностью v_γ. Это аналогично тому, что скачок интеграла типа Коши на границе Γ есть функция $\varphi(\xi)$. Функцию $\varphi(\xi)$ следует

понимать как плотность потенциала, если интеграл типа Коши понимать как потенциал.

Условие (8) есть условие того, что выражение (22) становится аналогом интеграла Коши (а не типа Коши). Именно в этом случае вторая строчка в правой части (22) обращается тождественно в нуль (интеграл Коши вне контура интегрирования обращается в нуль). Само условие (8) есть точный аналог условия Сохацкого–Племеля из теории аналитических функций, необходимое и достаточное для того, чтобы функция $\varphi(\xi)$ была граничным значением некоторой аналитической внутри D функции. В этом случае функция $f(z)$ на Γ совпадает с $\varphi(\xi)$. Редукция краевых задач к уравнениям с проектором на границе вполне аналогична методу сингулярных интегральных уравнений в теории краевых задач для аналитических функций.

Вместо квадратурных формул, используемых при численном решении сингулярных интегральных уравнений, в случае уравнений с проекторами используются разностные потенциалы с подходящей плотностью v_γ.

Задача

Определим функцию $u_{N^0} = P^-_{N^0\gamma} v_\gamma$:

$$P^-_{N^0\gamma} v_\gamma = \begin{cases} P^-_{N^-\gamma} v_\gamma, & n \in N^-, \\ -P^+_{N^+\gamma} v_\gamma, & n \in N^+, \end{cases} \tag{23}$$

аналогичную интегралу типа Коши (21) с интегрированием по контуру γ в направлении против часовой стрелки.

а) Показать, что в случае выполнения равенства (13) вторая строчка в правой части (23) есть функция, тождественно на N^+ обращающаяся в нуль.

б) Показать, что при любой функции v_γ справедливо равенство

$$P^-_{\gamma\gamma} v_\gamma - (-P^+_{\gamma\gamma} v_\gamma) = v_\gamma.$$

в) Показать, что $P^+_{\gamma\gamma}$ и $P^-_{\gamma\gamma}$ суть проекторы.

§ 4. Использование разностных потенциалов для приложений

Укажем некоторые приложения введенных выше разностных потенциалов.

1. Аппроксимация модифицированных потенциалов Кальдерона разностными. В § 2 мы свели задачу вычисления матрицы $C = [c_{ij}]$ и численного решения краевых задач к задаче приближенного вычисления потенциала Кальдерона $u_{\overline{D}} = P_{\overline{D}\Gamma} v_\Gamma$ с заданной плотностью

$$v_\Gamma = \begin{bmatrix} v^{(0)}(s) \\ v^{(1)}(s) \end{bmatrix}, \tag{1}$$

где $v^{(0)}(s)$, $v^{(1)}(s)$ — достаточно гладкие функции.

Оказывается, что при согласованном с (1) и указанном ниже задании разностной плотности v_γ разностный потенциал $u_{N^+} = P_{N^+\gamma}v_\gamma$ аппроксимирует потенциал Кальдерона $u_{\overline{D}} = P_{\overline{D}\Gamma}v_\Gamma$, так что значения $u_{\overline{D}}$ и ее первых производных в любой точке $x \in \overline{D}$ можно приближенно получить путем интерполяции функции u_{N^+} с сетки N^+. Опишем алгоритм задания v_γ. Пусть $v(x,y)$ $((x,y) \in \overline{D}^0)$ — некоторая функция, обращающаяся в нуль на границе квадрата D^0 и удовлетворяющая следующим условиям:

$$v\big|_\Gamma = v^{(0)}(s), \quad \partial v/\partial n\big|_\Gamma = v^{(1)}(s), \quad \Delta v\big|_\Gamma = 0, \quad \Gamma = \partial D, \tag{2}$$

где n — направление внутренней нормали к D. Заметим, что из условия $\Delta v\big|_\Gamma = 0$, зная $v^{(0)}(s)$ и $v^{(1)}(s)$, можно вычислить $\partial^2 v/\partial n^2\big|_\Gamma \equiv$ $\equiv v^{(2)}(s)$.

Пусть теперь ν — произвольная точка сеточной границы γ. Опустим из точки ν нормаль на Γ и значение длины дуги s в точке пересечения этой нормали с Γ обозначим s_ν. Расстояние от ν до точки s_ν на Γ обозначим ρ и будем считать положительным, если $\nu \in D$, и отрицательным, если $\nu \notin D$. Положим

$$v_\gamma = v^{(0)}(s_\nu) + \rho v^{(1)}(s_\nu) + \frac{\rho^2}{2} v^{(2)}(s_\nu), \quad \nu \in \gamma. \tag{3}$$

Эта формула задает сеточную плотность v_γ, для которой разностный потенциал $u_{N^+} = P_{N^+\gamma}v_\gamma$ аппроксимирует потенциал Кальдерона $u_D = P_{D\Gamma}v_\Gamma$ и позволяет вычислять потенциал u_D и его первые производные в произвольной точке с погрешностью, стремящейся к нулю при $h \to 0$. Мы не будем уточнять это утверждение А.А. Резника [16], а ограничимся рассмотрениями, объясняющими суть механизма аппроксимации.

Потенциал $u_{\overline{D}} = P_{\overline{D}\Gamma}v_\Gamma$ определен в § 1 как решение задачи

$$\Delta u = \begin{cases} 0, & x \in D, \\ \Delta v, & x \in D^0 \backslash D, \end{cases} \tag{4}$$

$$u\big|_{\partial D^0} = 0,$$

рассматриваемое на $\overline{D}$, $D \subset D^0$. Пусть плотность разностного потенциала задана не по формуле (3), а с помощью равенств

$$v_\nu = v(x_\nu, y_\nu), \quad \nu \in \gamma. \tag{5}$$

Обозначим эту плотность через $\widetilde{v}_\gamma$ в отличие от плотности v_γ, задаваемой формулой (3). Соответствующий разностный потенциал $\widetilde{u}_{N^+} = P_{N^+\gamma}v_\gamma$ в соответствии с § 3 совпадает на N^+ с решением

разностной задачи

$$\sum_{n\in N_m} a_{mn}\widetilde{u}_n = \begin{cases} 0, \quad m\in M^+, \\ \sum\limits_{n\in N_m} a_{mn}\, v(n_1h, n_2h), \quad m\in M^-, \end{cases} \tag{6}$$

$$u_n\big|_{\partial D^0} = 0.$$

Очевидно, что разностная задача (6) аппроксимирует задачу (4), так что естественно ожидать (так оно и есть), что $\widetilde{u}_{N^+}$ аппроксимирует $u_{\overline{D}}$. Но построение потенциала $\widetilde{u}_{N^+}$ нельзя осуществить, так как мы не знаем функции $v(x,y)$. Однако разностный потенциал зависит только от плотности $\widetilde{v}_\gamma$, которую мы приближенно можем вычислить по формуле Тейлора (3).

2. Перенос граничного условия с исходной границы на искусственную. Приведем пример задачи, требующей переноса граничного условия, а затем построим искусственное граничное условие, реализующее требуемый перенос.

Рассмотрим задачу

$$\begin{gathered} \sigma\,\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \sigma f(x,y,t), \quad 0\leqslant t\leqslant T, \\ u\big|_{t=0} = \varphi(x,y), \quad (x,y)\in D^0, \\ u\big|_{\partial D^0} = 0, \end{gathered} \tag{7}$$

где $f(x,y,t)$, $\varphi(x,y)$, $\psi(x,y)$, $\sigma = \sigma(x,y,t)$ — заданные функции, причем

$$\begin{gathered} \sigma(x,y,t)\equiv 0, \quad (x,y)\notin D\subset D^0, \\ \sigma(x,y,t) > \mathrm{const} > 0, \quad (x,y)\in D. \end{gathered} \tag{8}$$

Эта задача моделирует магнитное поле в проводнике D, окруженном вакуумом $D^0\backslash D$.

Относительно решения задачи (7) будем предполагать, что оно и его первые производные непрерывны на Г. Для вычисления решения сетку по переменным x, y возьмем ту же, что и в § 3. Аппроксимируем задачу (7) разностной схемой

$$\begin{gathered} \sigma_m\,\frac{u_m^{p+1} - u_m^p}{\tau} = \sum_{n\in N_m} a_{mn}u_n + \sigma_m f_m^p, \quad m\in M^0, \quad p = 1,2,\ldots, \\ u_m^0 = \varphi_m, \quad m\in M^0, \\ u_n^p\big|_{n\in\partial D^0} = 0. \end{gathered} \tag{9}$$

Допустим теперь, что решение $u(x,y,t)$ задачи (7) интересует нас только при $(x,y)\in D$. Соответственно решение задачи (9) требуется

вычислять только для $n \in N^+$. Допустим, что решение $u^p_{N^+}$ уже найдено при некотором p. В силу явной разностной схемы (9) мы можем вычислить u^{p+1}_n для всех $n \in M^+ \subset N^+$. Остается вычислить u^{p+1}_n в точках $N^+ \backslash M^+$, т. е. в точках множества $\gamma^- = \gamma \cap M^-$.

Обозначим подлежащую отысканию сеточную функцию u^{p+1}_n $(n \in \gamma^-)$ через $u^{p+1}_{\gamma^-}$. В соответствии с (9) функция $u^{p+1}_{\gamma^-}$ совпадает на γ^- $(\gamma^- \in N^-)$ с решением u_{N^-} следующей задачи:

$$\sum_{n \in N_m} a_{mn} u_n = 0, \quad m \in M^-, \tag{10}$$

$$u_n = u^{p+1}_n, \quad n \in \gamma^+ = \gamma \cap M^+, \tag{11}$$

$$u_n = 0, \quad n \in \partial D^0. \tag{12}$$

Условие (11) можно записать короче:

$$u_{\gamma^+} = u^{p+1}_{\gamma^+}, \tag{13}$$

где $u^{p+1}_{\gamma^+}$ — уже известная функция.

Очевидно, что u_{γ^-} есть такая сеточная функция, при которой функция

$$u_\gamma = \begin{cases} u^{p+1}_{\gamma^+} & \text{в точках } \gamma^+, \\ u_{\gamma^-} & \text{в точках } \gamma^- \end{cases} \tag{14}$$

допускает доопределение всюду в N^- до решения задачи (10), (12), (13). Поэтому u_γ должно удовлетворять условию (13) из § 3 или равносильному этому условию равенству (16) из § 3. Но функция z_{N^+}, определяемая решением задачи (2), (3) из § 3, совпадает с разностным потенциалом $u_{N^+} = P^+_{N^+\gamma} v_\gamma$ и зависит поэтому (и притом линейно) только от v_γ, или, что то же самое, от пары функций v_{γ^+}, v_{γ^-}.

Таким образом, функцию z_{γ^-}, совпадающую с z_{N^+} на γ^- $(\gamma^- \subset \subset N^+)$, можно записать в виде

$$z_{\gamma^-} = A_{\gamma^-\gamma^+} v_{\gamma^+} + B_{\gamma^-\gamma^-} v_{\gamma^-}, \tag{15}$$

где $A_{\gamma^-\gamma^+}$, $B_{\gamma^-\gamma^-}$ суть некоторые линейные операторы. Если плотность v_γ совпадает с функцией (14), то (15) примет вид

$$z_{\gamma^-} = A_{\gamma^-\gamma^+} u^{p+1}_{\gamma^+} + B_{\gamma^-\gamma^-} u_{\gamma^-}, \tag{16}$$

а условие (16) из § 3 примет вид

$$A_{\gamma^-\gamma^+} u^{p+1}_{\gamma^+} + B_{\gamma^-\gamma^-} u_{\gamma^-} = 0. \tag{17}$$

Это и есть граничное условие, которое получается при переносе условия (12) со множества точек сетки $n \in \partial D^0$ на искусственную

границу γ. Оно равносильно заменяет условие (12) и уравнение (10), вместе взятые.

Заметим, что оператор $B_{\gamma^-\gamma^-}$, действующий в пространстве сеточных функций v_{γ^-} размерности, равной числу $|\gamma^-|$ точек множества γ^-, является невырожденным и имеет обратный, так что (17) можно переписать в форме

$$u_{\gamma^-} = -(B_{\gamma^-\gamma^-})^{-1} A_{\gamma^-\gamma^+} u^{p+1}_{\gamma^+}. \tag{18}$$

Для доказательства существования оператора $B^{-1}_{\gamma^-\gamma^-}$ достаточно показать, что в случае $u^{p+1}_{\gamma^+} = 0_{\gamma^+}$ уравнение (17) имеет только тривиальное решение $u_{\gamma^-} = 0_{\gamma^-}$.

Пусть u_{γ^-} удовлетворяет (17) при $u^{p+1}_{\gamma^+} = 0$. Тогда функция (14) примет вид

$$u_\gamma = \begin{cases} 0 & \text{в точках} \quad \gamma^+, \\ u_{\gamma^-} & \text{в точках} \quad \gamma^- \end{cases}$$

и станет следом на γ решения однородного разностного уравнения

$$\sum_{n \in N_m} a_{mn} u_n = 0, \quad m \in M^-,$$

при условии

$$u_n\big|_{\partial D^0} = 0.$$

В силу принципа максимума максимум модуля этой функции достигается хотя бы в одной точке $N^- \backslash M^-$. Но множество $N^- \backslash M^-$ является объединением двух множеств — множества точек сетки, лежащих на границе квадрата ∂D^0, и множества γ^+. Но на первом из них $u_n = 0$ по условию $u_n|_{\partial D^0} = 0$. На втором также $u_n = 0$ в силу условия $u_n|_{\gamma^+} = u^{p+1}_{\gamma^+} = 0$. Поэтому $u_n \equiv 0$ $(n \in N^-)$ и, в частности, $u_{\gamma^-} = 0_{\gamma^-}$.

Для фактического вычисления $u^{p+1}_{\gamma^-} = u_{\gamma^-}$ по формуле (18) нужно перенумеровать точки множеств γ^+, γ^-. Тогда операторы $A_{\gamma^-\gamma^+}$, $B_{\gamma^-\gamma^-}$ запишутся с помощью матриц. Матрица $A_{\gamma^-\gamma^+}$ имеет $|\gamma^-|$ строк и $|\gamma^+|$ столбцов, а $B_{\gamma^-\gamma^-}$ — квадратная $|\gamma^-| \times |\gamma^-|$-матрица.

Элементом a_{ij} $(i = 1, 2, ..., |\gamma^-|,\ j = 1, 2, ..., |\gamma^+|)$ матрицы $A_{\gamma^-\gamma^+}$ является значение потенциала $P_{N^-\nu} v_\nu$ в точке с номером i на γ^-, если v_γ обращается в нуль всюду на γ, кроме точки с номером j на γ^+, где она равна единице. Элементом b_{ij} $(i, j = 1, 2, ..., |\gamma^-|)$ является значение этого же потенциала в той же точке с номером i множества γ^-, если v_γ обращается в нуль всюду, кроме точки с номером j на γ^-, где она равна единице. Таким образом, вычисление элементов a_{ij}, b_{ij} матриц сводится к вычислению разностных потенциалов.

После того как (18) записано с помощью матриц, значение $u_{\gamma^-}^{p+1} = u_{\gamma^-}$ вычисляется по формуле (18).

Подчеркнем, что матрица

$$C_{\gamma^-\gamma^-} = -B_{\gamma^-\gamma^-}^{-1} A_{\gamma^-\gamma^+},$$

входящая в искусственное граничное условие (18), не зависит от p и вычисляется один раз.

3. Идея декомпозиции задач с помощью метода разностных потенциалов. Изложим схему декомпозиции на примере задачи Дирихле для уравнения Лапласа в области D

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad (x, y) \in D, \tag{19}$$

$$u\big|_{\partial D} = \varphi(s). \tag{20}$$

Разобьем область D на две (несколько) подобласти, D_1 и D_2. Обозначим границы областей D_1, D_2 через Γ_1, Γ_2, а общую часть Γ_1 и Γ_2 через $\Gamma_{1,2}$. Объединение границ Γ_1 и Γ_2 обозначим Γ, так что $\partial D = \Gamma \backslash \Gamma_{1,2}$.

Уравнение Лапласа (19) в области D заменим уравнениями Лапласа

$$\frac{\partial^2 u^{(i)}}{\partial x^2} + \frac{\partial^2 u^{(i)}}{\partial y^2} = 0, \quad (x, y) \in D_i, \quad i = 1, 2, \tag{21}$$

с условиями согласования

$$(u^{(1)} - u^{(2)})\big|_{\Gamma_{1,2}} = 0, \quad \left(\frac{\partial u^{(1)}}{\partial n} - \frac{\partial u^{(2)}}{\partial n}\right)\bigg|_{\Gamma_{1,2}} = 0 \tag{22}$$

на общей части $\Gamma_{1,2}$ границы. Уравнение (19) и совокупность уравнений (21), (22) равносильны. Введем пространство V_{Γ_i} данных Коши на границе Γ_i. Введем вспомогательный квадрат $D_i^{(0)}$, содержащий D_i, и вспомогательную задачу

$$\begin{gathered} \frac{\partial^2 u^{(i)}}{\partial x^2} + \frac{\partial^2 u^{(i)}}{\partial y^2} = f^{(i)}(x, y), \quad (x, y) \in D_i^{(0)}, \\ u^{(i)}\big|_{\partial D_i^{(0)}} = 0, \quad i = 1, 2. \end{gathered} \tag{23}$$

Определим, как в § 1, потенциал Кальдерона для уравнения (21):

$$u_{D_i}^{(i)} = P_{D_i \Gamma_i}^{(i)} v_{\Gamma_i}, \quad i = 1, 2, \tag{24}$$

с плотностью $v_{\Gamma_i} \in V_{\Gamma_i}$ и выпишем условие на границе:

$$u_{\Gamma_i}^{(i)} = P_{\Gamma_i \Gamma_i}^{(i)} u_{\Gamma_i}^{(i)}, \tag{25}$$

которое равносильно соответствующему уравнению (21) в области D_i $(i = 1, 2)$.

Введем пространство V_Γ, состоящее из элементов вида $v_\Gamma = \begin{bmatrix} v_{\Gamma_1} \\ v_{\Gamma_2} \end{bmatrix}$ $(v_{\Gamma_i} \in V_{\Gamma_i},\ i = 1, 2)$. Введем также оператор $P_{\Gamma\Gamma}\colon V_\Gamma \to V_\Gamma$, положив

$$P_{\Gamma\Gamma} v_\Gamma = \begin{bmatrix} P^{(1)}_{\Gamma_1} v_{\Gamma_1} \\ P^{(2)}_{\Gamma_2} v_{\Gamma_2} \end{bmatrix}. \tag{26}$$

Совокупность уравнений (21) в областях D_1, D_2 равносильна, очевидно, совокупности уравнений (25), или

$$u_\Gamma = P_{\Gamma\Gamma}\, u_\Gamma, \tag{27}$$

где, очевидно, $P_{\Gamma\Gamma}$ — проектор, $P^2_{\Gamma\Gamma} = P_{\Gamma\Gamma}$.

Заменим условия (22) условиями более общего вида

$$(u^{(2)} - u^{(1)})\big|_{\Gamma_{1,2}} = \varphi^{(0)}_{\Gamma_{1,2}}, \quad \left(\frac{\partial u^{(2)}}{\partial n} - \frac{\partial u^{(1)}}{\partial n}\right)\bigg|_{\Gamma_{1,2}} = \varphi^{(1)}_{\Gamma_{1,2}}. \tag{28}$$

Совокупность условий (20), (28) запишем в виде

$$l u_\Gamma = \varphi_\Gamma. \tag{29}$$

Таким образом, исходная задача (19), (20) свелась к задаче вида (21), (29) в составной области $D = D_1 \cup D_2$, которая в свою очередь равносильна паре уравнений (27), (29) на границе $\Gamma = \Gamma_1 \cup \Gamma_2$. В этом и состоит предлагаемая декомпозиция исходной задачи.

Для численного решения системы граничных уравнений (27), (29) с проектором можно применить изложенную выше схему. Эта схема состоит в том, что плотность $v_\Gamma = \begin{bmatrix} v_{\Gamma_1} \\ v_{\Gamma_2} \end{bmatrix}$ заменяется плотностью $v^{(N)}_\Gamma$, зависящей от N значений в опорных точках на Γ, а для вычисления соответствующей матрицы (16) из § 2 потенциалы Кальдерона аппроксимируются разностными потенциалами на основе аппроксимации вспомогательных задач (23) их пятиточечными разностными аналогами на квадратной сетке с шагом $h_i (i = 1, 2)$.

Выгода от декомпозиции исходной задачи (19), (20) состоит в следующем.

а) В каждой области можно взять свой шаг h_i, отвечающий поведению решения в этой области.

Другими словами, можно совместить использование квадратной сетки с возможностью ее измельчения в нужных местах. (В методе конечных элементов возможность мельчить сетку достигалась переходом к нерегулярной треугольной сетке.)

б) Можно учесть произвольные граничные условия в криволинейной области, не прибегая к их разностной аппроксимации.

4. Использование многопроцессорных компьютеров. В основе алгоритмов МРП лежит многократное решение разностных вспомогательных задач, т. е. задач в простых областях (квадрат, круг, куб, шар) на регулярных сетках. Это допускает распараллеливание вычислений.

5. Библиографическая справка. Метод разностных потенциалов был предложен автором в его докторской диссертации в 1969 г. как аппарат численного решения разностных краевых задач. В 1976 г. (см.: Рябенький В. С. Об использовании метода внутренних граничных условий для вычислений //Тр. Всес. конф. по уравнениям с частными производными (январь 1976 г.).—М.: Изд-во МГУ, 1978.—С. 211–212) был предложен вариант МРП для численного решения дифференциальных краевых задач, не требующий разностной аппроксимации граничных условий.

В результате многолетних работ развился тот аспект МРП, основная задача которого изложена выше. Оказалось естественным трактовать этот аспект МРП как способ численного решения граничных уравнений с проекторами Кальдерона, которые были модифицированы с этой целью (Рябенький В. С. Проекторы Кальдерона// ДАН СССР.—1983.—Т. 270, № 2.—С. 288–291).

Работа по развитию МРП проводилась автором и работавшими вместе с ним в разные годы во время обучения на старших курсах и в аспирантуре Московского физико-технического института А.Я. Белянковым, Д.С. Каменецким, М.Ю. Лохановым, М.Н. Мишковым, А.А. Резником, И.Л. Софроновым, В.А. Торгашовым, А.М. Федоровским, С.В. Цынковым и другими. Участие в развитии МРП приняли Л.В. Бадьин, Р.И. Вейцман, А.В. Воронков, А.В. Забродин, Е.В. Зиновьев, М.И. Лазарев, В.В. Огнева, В.И. Турчанинов.

Основные результаты по МРП, полученные до 1987 г., нашли отражение в монографии [16]. Современное состояние МРП и его приложений отражено в монографии: Рябенький В. С. Метод разностных потенциалов и его приложения (которая готовится к выпуску в 2001 г.).

СПИСОК ЛИТЕРАТУРЫ

1. *Бабенко К.И.* Основы численного анализа. — М.: Наука, 1986.
2. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. — М.: Наука, 1987.
3. *Годунов С.К., Антонов А.Г. и др.* Гарантированная точность решения систем линейных уравнений в евклидовых пространствах. — Новосибирск: Наука, 1988.
4. *Годунов С.К., Забродин А.В. и др.* Численное решение многомерных задач газовой динамики. — М.: Наука, 1976.
5. *Годунов С.К., Рябенький В.С.* Введение в теорию разностных схем. — М.: Физматгиз, 1962.
6. *Годунов С.К., Рябенький В.С.* Разностные схемы. — М.: Наука, 1977.
7. *Дьяконов Е.Г.* Минимизация вычислительной работы. — М.: Наука, 1989.
8. *Жуков А.И.* Метод Фурье в вычислительной математике. — М.: Наука, 1992.
9. *Калиткин Н.Н.* Численные методы. — М.: Наука, 1978.
10. *Коробов Н.М.* Теоретико-числовые методы в прикладном анализе. — М.: Физматгиз, 1963.
11. *Магомедов К.М., Холодов А.С.* Сеточно-характеристический метод. — М.: Наука, 1988.
12. *Марчук Г.И.* Методы вычислительной математики. — М.: Наука, 1980.
13. *Михлин С.Г., Морозов Н.Ф., Паукшто М.В.* Граничные интегральные уравнения в теории упругости. — М.: Наука, 1993.
14. *Партон В.З., Перлин П.И.* Интегральные уравнения теории упругости. — М.: Наука, 1978.
15. *Петровский И.Г.* Лекции об уравнениях с частными производными. — М.: Гостехиздат, 1950.
16. *Рябенький В.С.* Метод разностных потенциалов для некоторых задач механики сплошных сред. — М.: Наука, 1987.
17. *Рябенький В.С., Филиппов А.А.* Об устойчивости разностных уравнений. — М.: Гостехиздат, 1956.
18. *Самарский А.А.* Теория разностных схем. — Изд. 2-е. — М.: Наука, 1983.
19. *Самарский А.А., Гулин А.В.* Численные методы. — М.: Наука, 1989.

20. *Самарский А.А., Николаев Е.С.* Методы решения сеточных уравнений. — М.: Наука, 1978.
21. *Соболев С.Л.* Уравнения математической физики. — Изд. 4-е. — М.: Наука, 1994.
22. *Соболев С.Л.* Введение в теорию кубатурных формул. — М.: Наука, 1973.
23. *Соболь И.М.* Численные методы Монте-Карло. — М.: Наука, 1973.
24. *Шайдуров В.В.* Многосеточные методы конечных элементов. — М.: Наука, 1989.
25. Высшая математика для инженерных специальностей: Компьютерное учебное пособие. Кн. 3, 4. — М.: РосНИИС, 1997.

ОГЛАВЛЕНИЕ

Предисловие 3
Введение 5
§ 1. Дискретизация 7
§ 2. Обусловленность 9
§ 3. Погрешность 11
§ 4. О методах вычисления 15

ЧАСТЬ I
ТАБЛИЧНОЕ ЗАДАНИЕ И ИНТЕРПОЛЯЦИЯ ФУНКЦИЙ. КВАДРАТУРЫ

ГЛАВА 1. **АЛГЕБРАИЧЕСКАЯ ИНТЕРПОЛЯЦИЯ** 19
§ 1. Существование и единственность интерполяционного многочлена 20
§ 2. Классическая кусочно многочленная интерполяция 27
§ 3. Кусочно многочленная гладкая интерполяция (сплайны) . . . 34
§ 4. Интерполяция функций двух переменных 42

ГЛАВА 2. **ТРИГОНОМЕТРИЧЕСКАЯ ИНТЕРПОЛЯЦИЯ** 44
§ 1. Интерполяция периодических функций 45
§ 2. Интерполяция функций на отрезке. Связь между алгебраической и тригонометрической интерполяциями 53

ГЛАВА 3. **ВЫЧИСЛЕНИЕ ОПРЕДЕЛЕННЫХ ИНТЕГРАЛОВ. КВАДРАТУРЫ** 59
§ 1. Квадратурные формулы трапеций и Симпсона 60
§ 2. Сочетание численных и аналитических методов при вычислении интегралов с особенностями 67
§ 3. Кратные интегралы 69

ЧАСТЬ II
СИСТЕМЫ СКАЛЯРНЫХ УРАВНЕНИЙ

ГЛАВА 4. **СИСТЕМЫ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ. МЕТОДЫ ОТЫСКАНИЯ ТОЧНОГО РЕШЕНИЯ** 72
§ 1. Формы записи совместных СЛАУ 73
§ 2. Нормы 77
§ 3. Обусловленность СЛАУ 82

§ 4. Методы исключения Гаусса 88
§ 5. Связь между задачей на минимум квадратичной функции и СЛАУ 96
§ 6. Метод сопряженных градиентов как метод точного решения СЛАУ 97
§ 7. Конечные ряды Фурье и запись точного решения разностного аналога задачи Дирихле для уравнения Пуассона 99

ГЛАВА 5. **МЕТОДЫ ПОСЛЕДОВАТЕЛЬНЫХ ПРИБЛИЖЕНИЙ (ИТЕРАЦИОННЫЕ МЕТОДЫ) РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ** 104

§ 1. Методы простых итераций 105
§ 2. Метод Чебышева и метод сопряженных градиентов 118

ГЛАВА 6. **ПЕРЕОПРЕДЕЛЕННЫЕ СЛАУ. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ** 120

§ 1. Примеры задач, приводящих к переопределенным СЛАУ . . . 120
§ 2. Переопределенные СЛАУ и обобщенные решения в общем случае 123

ГЛАВА 7. **ЧИСЛЕННОЕ РЕШЕНИЕ НЕЛИНЕЙНЫХ СКАЛЯРНЫХ УРАВНЕНИЙ И СИСТЕМ УРАВНЕНИЙ** 130

§ 1. Метод простых итераций 131
§ 2. Метод линеаризации Ньютона 135

ЧАСТЬ III

МЕТОД КОНЕЧНЫХ РАЗНОСТЕЙ ДЛЯ ЧИСЛЕННОГО РЕШЕНИЯ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

ГЛАВА 8. **ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ** 139

§ 1. Примеры разностных схем. Сходимость 139
§ 2. Аппроксимация дифференциальной краевой задачи разностной схемой 146
§ 3. Определение устойчивости разностной схемы. Сходимость как следствие аппроксимации и устойчивости 154
§ 4. Схемы Рунге–Кутта 161
§ 5. Методы решения краевых задач 165

ГЛАВА 9. **РАЗНОСТНЫЕ СХЕМЫ ДЛЯ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ** 169

§ 1. Основные определения и их иллюстрация 170
§ 2. Некоторые приемы построения аппроксимирующих разностных схем 183
§ 3. Спектральный признак устойчивости разностной задачи Коши 198
§ 4. Принцип замороженных коэффициентов 209
§ 5. Явные и неявные разностные схемы для уравнения теплопроводности 219

ГЛАВА 10. **ПОНЯТИЕ О РАЗРЫВНЫХ РЕШЕНИЯХ И СПОСОБАХ ИХ ВЫЧИСЛЕНИЯ** 221

§ 1. Дифференциальная формулировка интегрального закона сохранения 223
§ 2. Построение разностных схем 229

ГЛАВА 11. **РАЗНОСТНЫЕ МЕТОДЫ ДЛЯ ЭЛЛИПТИЧЕСКИХ ЗАДАЧ** 235

§ 1. Аппроксимация и устойчивость простейшей разностной схемы 237
§ 2. Понятие о методе конечных элементов 243
§ 3. Вычисление решений сеточных аналогов краевых задач 250
§ 4. Многосеточный метод Федоренко 251

ЧАСТЬ IV

МЕТОДЫ ГРАНИЧНЫХ УРАВНЕНИЙ ДЛЯ ЧИСЛЕННОГО РЕШЕНИЯ КРАЕВЫХ ЗАДАЧ

ГЛАВА 12. **ГРАНИЧНЫЕ ИНТЕГРАЛЬНЫЕ УРАВНЕНИЯ И МЕТОД ГРАНИЧНЫХ ЭЛЕМЕНТОВ ДЛЯ ИХ ЧИСЛЕННОГО РЕШЕНИЯ** . 256

§ 1. Способы редукции краевых задач к ГИУ 256
§ 2. Граничные элементы и дискретизация ГИУ 259
§ 3. Область применимости ГИУ для численного решения краевых задач 260

ГЛАВА 13. **ГРАНИЧНЫЕ УРАВНЕНИЯ С ПРОЕКТОРАМИ И МЕТОД РАЗНОСТНЫХ ПОТЕНЦИАЛОВ ДЛЯ ИХ ЧИСЛЕННОГО РЕШЕНИЯ** 261

§ 1. Пример ГУРП для эллиптического уравнения второго порядка 263
§ 2. Устойчивость ГУРП и проблема их дискретизации 267
§ 3. Определение и вычисление разностных потенциалов 276
§ 4. Использование разностных потенциалов для приложений . . . 282

Список литературы 290

Учебное издание

РЯБЕНЬКИЙ Виктор Соломонович

ВВЕДЕНИЕ В ВЫЧИСЛИТЕЛЬНУЮ МАТЕМАТИКУ

Редактор *Е.Ю. Ходан*
Оригинал-макет *Н.Л. Ивановой*

ЛР № 071930 от 06.07.99. Подписано в печать 25.10.2000.
Формат 60×90/16. Бумага типографская № 1. Печать офсетная.
Усл. печ. л. 18,5. Уч.-изд. л. 20,9. Тираж 3000 экз. Заказ №

Издательская фирма "Физико-математическая литература"
МАИК "Наука/Интерпериодика"
117864 Москва, ул. Профсоюзная, 90

Отпечатано с готовых диапозитивов в Московской типографии № 6
Министерства РФ по делам печати, телерадиовещания
и средств массовых коммуникаций
109088 Москва, Ж-88, Южнопортовая ул., 24

